The Quest to Replace Passwords Revisited – Rating Authentication Schemes

V. Zimmermann[1], N. Gerber[2], M. Kleboth[1], A. von Preuschen[1], K. Schmidt[1]
and P. Mayer[2]

[1]FAI - Work and Engineering Psychology - Technische Universität Darmstadt
[2]SECUSO - Security, Usability, and Society, Karlsruhe Institute for Technology
e-mail: zimmermann@psychologie.tu-darmstadt.de

Abstract

Six years ago Bonneau et al. (2012) proposed a framework to comparatively evaluate authentication schemes. They applied their framework to 35 different authentication schemes to identify alternatives to the ubiquitous text password. However, in their work no sole authentication scheme proved to be suitable for every application scenario, hence the quest to replace passwords has not yet been solved. This paper revisits the rating process and describes the application of an extended version of the original framework to an additional 40 authentication schemes identified in a literature review. All schemes were rated in terms of 25 objective features assigned to the three main criteria usability, deployability, and security. The rating process and results are presented along with a discussion of the benefits and pitfalls of the rating process. Our goal thereby is not to claim victory over text passwords, but to help decision makers in identifying suitable authentication schemes for their specific application scenario. The results were also made publicly available in an authentication choice support system named ACCESS to foster the further extension of the knowledge base and future development of the rating process.

Keywords

Authentication Scheme, Password, Rating, ACCESS

1. Introduction

Authentication has long become an integral part of daily life. Every single authentication process provides access to private data like emails, account data, personal documents, or photos. A loss thereof to an unauthorized third party can thus have a huge impact on private life or businesses. The password as an authentication scheme still is ubiquitous. Although it is often used for various reasons such as low technical requirements or habit, the security of the scheme very much depends on the end user. With every new user account and every new password cognitive load is increasing so that usability is often preferred over security by users: For example, users often choose the same password across accounts, keep an insecurely stored written record or choose unsecure dictionary passwords (e.g., Adams *et al.* 1997, Johnson and Grawemeyer, 2011, Wash *et al.*, 2016).

To mitigate the issues associated with text passwords, many alternative schemes have been developed including biometric or token-based schemes. Bonneau *et al.* (2012) compared these to the text password across a variety of features and, surprisingly, found that replacing the password was not as easy as imagined. None of the analyzed schemes received high scores in all of the three evaluated categories usability, deployability, and security. Still, the comparison has proven to be very helpful in identifying authentication schemes best-suited for a certain purpose or certain requirements in research and practice alike. Thus, the initial work by Bonneau *et al.* (2012) serves as a basis for the evaluation of further authentication schemes. To realize an even more objective evaluation with an increased differentiation between authentication schemes additional sub features have been formulated by Mayer *et al.* (2016). The sub features were formulated as partially exclusive axioms to clearly allocate a scheme to a certain class of features.

Application of the evaluation framework by Bonneau *et al.* (2012) and the refinement of Mayer *et al.* (2016) facilitates an objective comparison between authentication schemes and allows for the selection of schemes fulfilling specific application requirements. However, while their results demonstrate the suitability of the rating for researchers and practitioners, the coverage of authentication schemes by their work is still very limited. Mayer *et al.* (2016) applied their finer-grained ratings only to the original data set from Bonneau *et al.* (2012) and an additional ten schemes. Compared to these 45 schemes, a far greater number of schemes have been proposed in the literature and decision-makers in research and practice would greatly benefit from an update and extension of the data set to choose suitable authentication schemes from. In order to advance the diversity of authentication scheme in the rated pool, this paper describes the process and results of a rating of 40 additional authentication schemes identified in the literature. The core contributions of this work are three-fold:

1. The pool of authentication schemes rated using the same methodology is significantly extended from 45 to 85. Thereby, not only the number, but also the diversity in the pool of available schemes is increased. This extension offers decision makers a greater selection when choosing appropriate authentication schemes for their specific application scenarios.

2. The ratings are integrated into the free, online authentication choice support system ACCESS (Renaud et al., 2014; SECUSO, 2016) so that practitioners and researchers can easily benefit from our results.

3. The advantages and pitfalls of the rating process are discussed to support others in the future rating of authentication schemes and to provide a starting point for solving ambiguous results within the community.

The remainder of the paper is structured as follows. Section 2 describes the methodology of the rating process. Section 3 presents exemplarily the rating results. Due to space constraints, the complete rating results are made available within ACCESS (c.f. contribution 2). In section 4, use cases as well as advantages and limitations of the rating process are discussed. Finally, section 5 concludes the paper.

2. Method

One of the primary goals of this research was to supplement and update the original rating of authentication schemes by Bonneau *et al.* (2012). Further, the aim was to increase the level of detail and thereby the usefulness of the rating for researchers and practitioners. To that end, a literature search via Google Scholar was conducted which revealed a total of 164 relevant publications dealing with authentication schemes. All publications addressed or evaluated the user interaction with or perception of the authentication schemes. Papers only describing technical aspects or algorithms were not considered. From the analysis 40 authentication schemes which were not already included in the rating by Bonneau *et al.* (2012) could be extracted. Even though all schemes were extracted from research papers, a significant number of these schemes are actually used in practice, e.g., Challenge Questions, Face Recognition, Passphrases, and Google's Android Pattern Unlock.

Our second step was to rate the schemes according to the 63 sub features specified by Mayer *et al.* (2016) and shown in Appendix B. These were derived from the original 25 features of authentication schemes as defined by Bonneau *et al.* (2012). The sub features are extensions of the original features and provide a more detailed way to evaluate authentication schemes. For example, the feature "Accessible" is split into the three sub features "Accessible with Read/Write-Impairments", "Accessible with Visual Impairments" and "Accessible with Physical Impairments". They are also partially exclusive in that a scheme can only fulfil one of the sub features but not two at the same time. This allows for the allocation of schemes to distinctive classes. An example for this is the feature "Proprietary" with the sub features "Proprietary" or "Non-Proprietary".

The rating process was structured as follows: similar to Bonneau *et al.* (2012) three of the authors each rated a subset of the 40 identified authentication schemes in terms of every sub feature. Any arising questions or problems were discussed within the research group including an additional three independent researchers. Whenever possible, the rating was based on the description of the scheme or other data provided by the authors in the original publication. Where the original publication was not available or sufficient, e.g. where the scheme was only described in a review paper, additional literature describing the scheme was considered. In case a publication did not provide any specifics regarding a criterion, e.g., because the scheme was presented only on a conceptual level, the rating was logically derived from the description of the scheme. For example, even though some descriptions of biometric schemes did not actually state the number of secrets to remember to rate the feature "Memorywise-Effortless" the information was logically derived from the conceptual approach which is based on detecting biometric features that users carry with them naturally and do not have to remember. All ratings were conducted for using the authentication scheme with a PC or laptop.

In general, the ratings of authentication schemes widely used in various forms and without an identifiable "original" publication such as the fingerprint scheme or different password schemes were based on the *concept* of the scheme, rather than the specifics of a *certain implementation*. The idea behind this approach was that a scheme should not be excluded by a decision-maker beforehand due to a low rating based on a single implementation. A researcher or practitioner deciding to use such a scheme could easily adapt certain aspects of an implementation according to his or her context of use. An example is setting a limit to the number of login attempts allowed before temporarily blocking an account, which affects the rating of the feature "Resilient-to-Throttled-Guessing". To preserve internal consistency, all new schemes were also compared to the ones that had already been rated by Bonneau *et al.* (2012) and Mayer *et al.* (2016) thus giving similar authentication schemes identical ratings. Examples include the already rated "Iris Scan" that shares features with the newly added "Retina Scan".

3. Results

Due to space constraints, the rating results will be presented exemplarily for three authentication schemes and two usability, deployability and security features each. The three exemplary schemes are Retina Scan, Google's Android Pattern Unlock and the scheme Déjà Vu as proposed Dhamija and Perrig (2000). The complete rating results can be accessed online and via ACCESS (see Appendix A).

Retina Scan is a biometric authentication scheme that identifies the user by his/her unique patterns on the retina blood vessels (Figure 1a). The patterns are detected optically by casting an unperceived beam of low-energy infrared light into the user's eye and measuring the absorption levels of light. In general, an appropriate scanner is required to perform the authentication. The Retina Scan is different from the Iris Scan where near infrared images of the iris are used for authentication. Similar to the Finger Print the Retina Scan is a general concept with a variety of implementations.

Android Pattern Unlock is a recall-based graphical authentication scheme mainly used on mobile phones. To authenticate the user draws a memorized path visiting up to nine dots on a 3x3 grid. Each dot can only be visited once (Figure 1b).

Déjà Vu is a recognition-based graphical scheme. The user memorizes a portfolio of pictures, which are algorithmically generated from random seeds (see Figure 1c). During authentication the user is provided with a challenge set that contains some of the images from his/her portfolio as well as a number of distractors. The user's task is to identify the previously chosen images from the challenge set. In contrast to the more general concept of retina scans, Déja Vu is described as a specific authentication scheme with an original publication and implementation details.

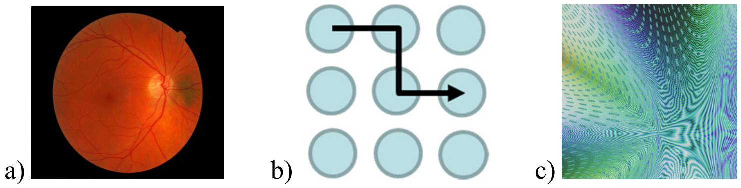

Figure 1: a) Retinal blood vessels used for Retina Scans, b) Android Pattern Unlock, c) Random art "Déjà Vu" picture from www.random-art.org

Usability. In the usability category Bonneau's feature “Memorywise-effortless” is split into the three exclusive sub features: “No-Secret-to-Remember”, “One-Secret-to-Remember” and “More-than-One-Secret-to-Remember”. As Retina Scans solely rely on measurable characteristics of the user, they are assigned the feature “No-Secret-to-Remember”. “One-Secret-to-Remember” is mainly provided by systems like single sign-on services that require only a single secret to gain access to multiple different systems. This is not the case for Déjà Vu and Android Pattern Unlock, which require the user to create a new, individual secret for each verifier and consequently were rated "More-than-one-Secret-to-Remember". Moreover, according to the original publication the scheme Déjà Vu requires the user to be able to recognize multiple pictures per login.

The schemes Déjà Vu and Android Pattern lock were rated “easy-recovery-from loss” as forgotten or stolen secrets could easily be replaced by new ones without having to overcome unreasonable burdens, e.g. by sending a recovery link via email. In contrast, Retina Scan was rated as “no-easy-recovery-from-loss” as a compromised account or a physical inability to further use the scheme results in having to replace the scheme with an alternative one.

Deployability. In terms of deployability the feature “Negligible-Cost-per-User” is considered. For retinal scans the standard cameras in laptops and smartphones are not feasible as measuring instruments. Thus, the user or service provider has to purchase additional scanning devices which results in high acquisition costs. Accordingly, Retina Scans were rated to have "Non-Negligible-cost-per-User". The scheme Déja Vu can be used with any standard PC and browser so that no additional devices have to be acquired. And even though Déjà Vu requires the verifier to store multiple seed values for the generated pictures in a secure manner, the resulting costs were considered to be negligible which resulted in the "Negligible-Cost-per-User" rating. From a technical perspective Android Pattern Unlock also requires negligible-cost-per-user for implementation. It can theoretically be used with any mobile phone, PCs with a touch screen and standard PCs using a mouse to draw the path in the grid. Still, the scheme has been developed and patented by Google (Google Inc., 2011). As we were not able to quantify potential license fees, e.g., for commercial purposes, we assumed “negligible-cost-per-user” but marked with a “?”.

Another deployability feature is the maturity of the schemes. Google’s Android Pattern Unlock is well studied in the literature and widely used in large number of Android mobile phones. Similarly, Retina Scans have been studied in academia and are used in practice, e.g., by government agencies, for medical purposes and ATM identity verification. Both schemes were thus granted all three sub features “adopted-in-academics”, “adopted-repeatedly” and “adopted-beyond-academics”. Déjà Vu has been proposed in the literature, but we are not aware of an application outside academia. The schemes was thus granted “adopted-in-academics” only.

Security. A scheme is considered "Non-Resilient-to-Phishing" if a potential attacker only needs to feign the identity of the verifier to obtain authentication credentials from users. More sophisticated methods of phishing, for example schemes that require the attacker to pose as a user and as a verifier are not considered in the definition by Mayer *et al.* (2016). As the Retina Scan and Android Pattern Unlock only involve a static characteristic, namely the unique patterns on the retina blood vessels and the string resulting from the path on the grid respectively, and an attacker only needs to pose as a verifier we rated the method as "Non-Resilient-to-Phishing". In contrast the Déjà Vu scheme is rated "Resilient-to-Phishing", as the attacker first needs to pose as the user to obtain the user specific challenge, which he or she then needs to present when posing as verifier. Additionally, it is not possible to obtain the entire user portfolio within one trial, since only a subset of chosen pictures is presented in each challenge set.

As the schemes Déjà Vu and Android Pattern Unlock require an active user input, they cannot be executed without the user's consent and are thus granted the feature "requiring-explicit-consent". The scheme Retina Scan requires a certain scanning device and an exact positioning of the user. It is thus unlikely that the authentication takes place without the user noticing. The scheme was therefore rated "requiring-explicit-consent" as well. Still, it is possible to track certain other biometrics, e.g., capturing the face with a camera or the keystroke dynamics while typing, without the user taking notice which would result in the rating "non-requiring-explicit-consent".

4. Discussion

The following section presents examples for the application of the rating by researchers and practitioners and discusses benefits and limitations of the rating process in its current form. Further, an outlook on the application of the rating within ACCESS (Mayer*, et al.* 2016) is provided.

4.1. Application of the Rating by Researchers

The results of the rating process can be useful for authentication research as they allow researchers to quickly identify appropriate authentication schemes for study purposes or software applications developed within a research project. It further allows a thorough comparison of newly developed authentication schemes with a variety of existing approaches on the three categories usability, deployability and security. One practical example for the use of the rating is a project on user-friendly authentication and encryption within the Centre for Research in Security and Privacy (CRISP). Within the project certain limitations for the choice of the authentication scheme exist, e.g., it should be cost-free for the user, deployable in web browsers, and users should not need to carry additional items for authentication. Further, even though it is impossible to determine an absolute security value, the authentication scheme (and thus the encrypted communication) should be resistant to a variety of attacks and relatively secure compared to other authentication schemes. First, the rating process described here allowed for excluding authentication schemes that did not meet the criteria set in the project and rank others in terms of the remaining objective security, usability and deployability features. Second, the rating was used to identify the best performing schemes out of five different categories, such as knowledge-based and biometric schemes. The resulting schemes were analysed in terms of user perceptions in a laboratory study revealing three schemes preferred by the participants. These three schemes will now be evaluated against each other, e.g., within mock-ups, to identify the most suitable one for this use case.

4.2. Application of the Rating by Practitioners

Practitioners may use the results of the rating for similar purposes as researchers, e.g. for study purposes or for comparing own with existing approaches. Apart from that, the rating may support practitioners in identifying an appropriate authentication scheme for their service, web application, or product. It provides an overview over a range of existing schemes and, similar to the research example described above, allows excluding schemes that do not meet the requirements given by the product or the target user group.

4.3. Benefits and Limitations of the Rating Process

As described above, the rating process provides a number of benefits for researchers and practitioners alike: support in the choice of an existing authentication scheme for one's own application or study, a comparison of new schemes with existing ones, and requirement- as well as context-based ratings of authentication schemes. Still, the rating process in its current form and the results described here suffer from several shortcomings that should be acknowledged and addressed in the future:

First, the rating process was based on the literature available to us. Some schemes, e.g., "Marbles" (von Zezschwitz *et al.,* 2013) which is an authentication scheme originally designed for smart phones with the aim to avoid smudge attacks were only described in a few papers or on a conceptual level. In particular, some details and technical information necessary for the rating were not available, so that the rating had to be based on similar schemes and/or logically derived from the conceptual approach. For the future the rating would thus benefit from being checked for correctness by the developers of the rated authentication schemes that are experts for their work. Other schemes, however, were described in many papers and in many different forms or implementations. One example is keystroke dynamics, where various implementations and service providers exist. In this case, the broader concept of authentication using keystroke dynamics independent from a single implementation was rated. In cases where this was not possible, we searched for review papers or a "common" way of implementation. Still, for future work it might be beneficial to rate and name different implementations separately and include the reference to the developers of that implementation.

Second, the rating was conducted at one point in time and with certain search terms and thus does not claim to cover an exhaustive list of existing authentication schemes. Besides, it is possible that schemes have been developed and improved further or that schemes are not available any more. To provide a valuable and actual resource for researchers and practitioners it would therefore be beneficial if the database would be regularly checked and updated by members of the community.

One way to allow for the checking of the rating scores by the developers of authentication schemes and the regular updating of the database by the community is provided by ACCESS, the authentication choice system that is presented in more detail in section 4.4.

4.4. Outlook

The rating process described in this paper was purposefully based on the criteria used within ACCESS of which a first version has already been presented in Renaud *et al.* (2014) and implemented by Mayer *et al.* (2016). A second version has now been released (Mayer *et al*, 2018). ACCESS supports authentication researchers and practitioners in providing information on the included authentication schemes (information module) and showing the five most suitable schemes given the weighting or exclusion of certain sub features according to the usage scenario (decision support module). The third feature, the discussion module, allows for updating and extending the knowledge base with additional authentication schemes.

For researchers and practitioners alike, the major benefit of ACCESS is that it presents the results of the rating process described above in a comprehensive and easily manageable form. All schemes are briefly described so that also practitioners not familiar with the schemes are provided with basic information. The decision support module allows for an easy individualization of the decision process for an authentication scheme. For example, developers of an online web service might assign high priority to browser-compatibility and likely aim to exclude costly schemes. They could easily arrange these features according to their preference in a drag and drop menu and be provided with a list of the best performing schemes given their individual use case.

To be consistent with the ACCESS knowledge base, the aforementioned ratings were used to generate equivalence classes for all authentication schemes similar to Mayer *et al.* (2016). The final step has been the transfer of the rating results presented here to the ACCESS database, thereby increasing the number of included authentication schemes from 45 to 85. With the provision of our results within ACCESS we hope to allow a large number of researchers and practitioners to benefit from our work. Further, we hope to thereby encourage other members of the community to add further schemes to the platform and participate in discussing and solving potential ambiguities in the rating process.

5. Conclusion

This paper describes the rating process of 40 authentication schemes in terms of the three categories usability, security and deployability based on the framework introduced by Bonneau *et al.* (2012) and refined by Mayer *et al.* (2016). The rating offers researchers as well as practitioners an aid in the choice of appropriate authentication schemes for their specific application scenarios and allows comparisons with newly developed schemes. To make the results easily available for the community, the rating results have been included into the knowledge base of ACCESSv2 (SECUSO, 2016), an authentication choice support system that allows the requirement-based rating of the authentication schemes. ACCESS also enables regular updating and correction of the data by the community and the developers of authentication schemes. Finally, the advantages and pitfalls of the rating process were discussed to support others in the future rating of authentication schemes and to provide a starting point for solving ambiguous results within the community.

6. Acknowledgement

The research reported in this paper was supported by the German Federal Ministry of Education and Research (BMBF) and by the Hessian Ministry of Science and the Arts within CRISP (www.crisp-da.de/). This work was further supported by the German Federal Ministry of Education and Research in the Competence Center for Applied Security Technology (KASTEL).

7. References

Adams, A., Sasse, M. A., and Lunt, P. (1997), “Making passwords secure and usable.”, in *People and Computers XII*, pp1-19, Springer, London.

Bonneau, J., Herley, C., van Oorschot, P.C. and Stajano, F. (2012), “The Quest to Replace Passwords: A Framework for Comparative Evaluation of Web Authentication Schemes”, in *Proceedings of the IEEE Symposium on Security and Privacy 2012*, pp553-567, IEEE.

Dhamija, R. and Perrig, A. (2000, August), “Deja Vu-A User Study: Using Images for Authentication”, in *Proceedings of the 9th USENIX Security Symposium,* Usenix.

Google Inc. (2011). *US Patent No. 20110283241,* Touch Gesture Actions From A Device's Lock Screen. Washington, D.C.: US Patent & Trademark Office.

Grawemeyer, B. and Johnson, H. (2011), “Using and managing multiple passwords: A week to a view. “, *Interact. Comput. 23*, pp256–267, doi: 10.1016/j.intcom.2011.03.007.

Mayer, P., Neumann, S., Storck, D. and Volkamer, M. (2016), “Supporting Decision Makers in Choosing Suitable Authentication Schemes”, in *Proceedings of the Tenth International Symposium on Human Aspects of Information Security & Assurance (HAISA 2016)*, pp67-77.

Mayer, P., Stumpf, P., Weber, T., Volkamer, M. (2018), ACCESSv2: A Collaborative Authentication Research and Decision Support Platform, Who Are You?! Adventures in Authentication.

Renaud, K., Volkamer, M. and Maguire, J. (2014, June), “ACCESS: Describing and Contrasting”, in *International Conference on Human Aspects of Information Security, Privacy, and Trust*, pp183-194, Springer, Cham.

Von Zezschwitz, E., Koslow, A., De Luca, A. and Hussmann, H. (2013, March), “Making graphic-based authentication secure against smudge attacks”, in *Proceedings of the 2013 international conference on Intelligent user interfaces*, pp277-286, ACM.

Wash, R., Rader, E., Berman, R. and Wellmer, Z. (2016, June), “Understanding password choices: How frequently entered passwords are re-used across websites”, in *Symposium on Usable Privacy and Security (SOUPS)*, pp175-188, Usenix.

Appendix A: Online-Appendix

The complete results and a description of the rated authentication schemes and rating features can be accessed with the following link: http://www.arbing.psychologie.tu-darmstadt.de/home/forschung_4/forschungsergebnisse_fai.de.jsp

The rating results are further integrated in ACCESS: https://access.secuso.org/

Appendix B: Rating Features

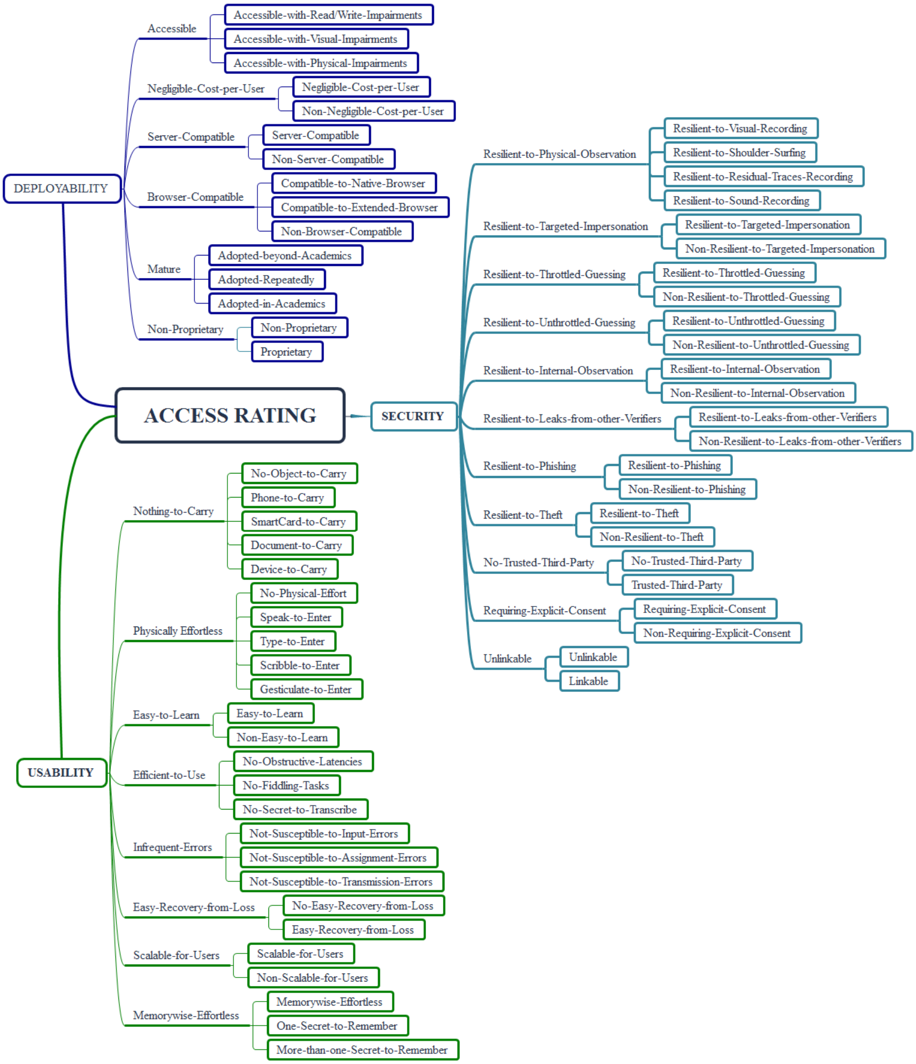
ACCESS RATING
DEPLOYABILITY
Accessible
Accessible-with-Read/Write-Impairments
Accessible-with-Visual-Impairments
Accessible-with-Physical-Impairments
Negligible-Cost-per-User
Negligible-Cost-per-User
Non-Negligible-Cost-per-User
Server-Compatible
Server-Compatible
Non-Server-Compatible
Browser-Compatible
Compatible-to-Native-Browser
Compatible-to-Extended-Browser
Non-Browser-Compatible
Mature
Adopted-beyond-Academics
Adopted-Repeatedly
Adopted-in-Academics
Non-Proprietary
Non-Proprietary
Proprietary
SECURITY
Resilient-to-Physical-Observation
Resilient-to-Visual-Recording
Resilient-to-Shoulder-Surfing
Resilient-to-Residual-Traces-Recording
Resilient-to-Sound-Recording
Resilient-to-Targeted-Impersonation
Resilient-to-Targeted-Impersonation
Non-Resilient-to-Targeted-Impersonation
Resilient-to-Throttled-Guessing
Resilient-to-Throttled-Guessing
Non-Resilient-to-Throttled-Guessing
Resilient-to-Unthrottled-Guessing
Resilient-to-Unthrottled-Guessing
Non-Resilient-to-Unthrottled-Guessing
Resilient-to-Internal-Observation
Resilient-to-Internal-Observation
Non-Resilient-to-Internal-Observation
Resilient-to-Leaks-from-other-Verifiers
Resilient-to-Leaks-from-other-Verifiers
Non-Resilient-to-Leaks-from-other-Verifiers
Resilient-to-Phishing
Resilient-to-Phishing
Non-Resilient-to-Phishing
Resilient-to-Theft
Resilient-to-Theft
Non-Resilient-to-Theft
No-Trusted-Third-Party
No-Trusted-Third-Party
Trusted-Third-Party
Requiring-Explicit-Consent
Requiring-Explicit-Consent
Non-Requiring-Explicit-Consent
Unlinkable
Unlinkable
Linkable
USABILITY
Nothing-to-Carry
No-Object-to-Carry
Phone-to-Carry
SmartCard-to-Carry
Document-to-Carry
Device-to-Carry
Physically Effortless
No-Physical-Effort
Speak-to-Enter
Type-to-Enter
Scribble-to-Enter
Gesticulate-to-Enter
Easy-to-Learn
Easy-to-Learn
Non-Easy-to-Learn
Efficient-to-Use
No-Obstructive-Latencies
No-Fiddling-Tasks
No-Secret-to-Transcribe
Infrequent-Errors
Not-Susceptible-to-Input-Errors
Not-Susceptible-to-Assignment-Errors
Not-Susceptible-to-Transmission-Errors
Easy-Recovery-from-Loss
No-Easy-Recovery-from-Loss
Easy-Recovery-from-Loss
Scalable-for-Users
Scalable-for-Users
Non-Scalable-for-Users
Memorywise-Effortless
Memorywise-Effortless
One-Secret-to-Remember
More-than-one-Secret-to-Remember

Figure 2: Categories, features (Bonneau *et al.*, 2012) and sub features (Mayer *et al.*, 2016) applied in the rating process.