Confidentiality of Electronic Documents in the Civil Registry of Brazil

Wellington Fernandes Silvano[1], **Frederico Schardong**[1], **Lucas Mayr**[1], **Eduardo Perottoni**[1], **Ricardo Custódio**[1]

LabSEC, Federal University of Santa Catarina (UFSC), Santa Catarina - Brazil

Abstract. *This work presents a cryptographic protocol for securing electronic documents in the Civil Registry of Brazil. The protocol is innovative because the document is encrypted and stored in the cloud but can be easily recovered by the owner if she has the access credentials, that is, a simple password and the owner identifier. The main objective is to keep the management of credentials by users as simple as possible while preventing anyone from accessing the stored document. The protocol was implemented using advanced electronic signature services of PDF standard. The implementation separates the public parts of the document, storing, in encrypted form, exclusively the private data and cryptographic artifacts.*
NOTE: This is still a work in progress.

1. Introduction

The security of personal data is a central concern in the digital era and it is crucial to address this issue in various domains, including institutions responsible for maintaining records, such as registry offices. Registry offices bear the responsibility of managing a vast amount of official documentation, ensuring the authenticity, integrity, and confidentiality of these documents in compliance with laws and regulations.

Civil Registry is subject to specific laws and regulations that require adequate protection of entrusted documents and storage, particularly given its importance in ensuring the authenticity and integrity of the documents that comprise our legal identity. The Public Records Act [Brasil 1973] and other regulations governing registries' activities emphasize the importance of adequate record preservation, protection, and accessibility. General law and data protection (LGPD) [Brasil 2018] establishes principles and guidelines for personal data processing, requiring institutions to ensure the security and privacy of personal data.

The digitization of the process of requesting and receiving official documents from Registrar Offices has the potential to improve the efficiency and accessibility of official document retrieval, in line with the LGPD's principle of free access. Nonetheless, many concerns have been raised about the confidentiality of electronic documents online, as opposed to the previously paper-based system. Although obtaining documents online may save citizens time and money, one must question whether they are in a more secure position regarding privacy. Therefore, systems must guarantee information confidentiality to mitigate electronic system risks, such as illicit access to personal data.

In this context, developing confidential models is critical to ensuring the confidentiality of personal data and documents. Such models must ensure personal data's integrity, authenticity, and privacy while meeting applicable legal and regulatory standards. This

paper describes a protocol for online distributing official documents issued by Civil Registrars. It is based on exploiting a characteristic of the documents created by these entities to extract and securely store personal data. Our scheme conceals private information to prevent attackers from accessing citizen data at rest while allowing citizens to retrieve their documents.

Given that civil records pertain to the entire population, it is necessary to facilitate access to the documents they generate. Not only does our proposal address data security, but also its accessibility to citizens by requiring *minimal hardware* on their end. The concept of *minimal hardware* emerges as a method that tries to foster inclusiveness by ensuring that devices and systems are created with all users, including those with low digital resources and skills, in mind. Digital inclusion is critical to ensuring that all people, regardless of capabilities or financial resources, can fully participate in the digital society.

In addition to cryptographic protocols, the physical-to-digital transformation and its peculiarities must be considered when converting documents to the digital realm. In Provision 63 [CNJ 2017], for instance, the issuance of official civil documents follows predefined rules. Civil Registry offices can only legally issue official papers, *e.g.*, birth and marriage certificates, if printed on a ratified security paper [CNJ 2012, CNJ 2011]. Printing is restricted to registry offices, increasing confidence in the document's legitimacy.

Contributions. In this paper, we present a cryptographic protocol associated with a new access control policy that focuses on protecting personal data and the confidentiality of documents while adhering to all legal obligations and responsibilities. Additionally, we optimize the storage of these documents in PDF format, enabling the storage of hundreds of millions of encrypted documents on consumer-grade hardware. This study aims to improve availability and privacy while maintaining online records' integrity and confidentiality.

2. Fundamentals

This session aims to present foundational concepts discussed in this article, such as the need for encryption in external Storage and access policies, the exploration of electronic documents focused on ISO 32000 and PadES, the general data protection law with a focus on confidentiality, and the current normative of the Civil Registry.

2.1. Encryption, External Storage, and Access Policies.

The storage of data on an external platform raises concerns regarding the security and privacy of the data [Banerjee et al. 2017]. Users are required to place trust in the credibility of the server [Raja and Ramakrishnan 2020]. To ensure the confidentiality of the data, techniques such as encryption have been employed to protect data [Mohanty et al. 2016, Khan et al. 2016]. Data encryption involves the strict encryption of sensitive data or files stored on the server-side, rendering the encrypted data difficult to comprehend even in the event of a leak, thereby enhancing data security [Feng et al. 2017, Wu et al. 2012]. Encryption serves as a significant means to address the issue of data confidentiality in external server [Liu et al. 2017, Li et al. 2022].

Relying solely on login access systems for accessing information stored on external servers poses significant challenges.This situation is not particularly attractive from a security perspective. In the event of server compromise, for instance, as a result of software vulnerability exploitation, the potential for information theft is immense. Additionally,

there is always a risk of "internal attacks" where an individual with server access steals and leaks information, for example, for economic gain [Goyal et al. 2006].

2.2. Electronic Documents: ISO 32000 and PAdES

The Portable Document Format (PDF) is widely used for electronic documents in civil registration, although it is not mandatory. The ISO 32000 standard defines the technical specifications and requirements for the creation, structure, processing, display, and exchange of PDF documents. It covers elements such as markup, objects, fonts, images, annotations, forms, digital signatures, and encryption [ISO 2008]. For digital signatures in PDF, the PAdES standard is widely adopted, ensuring integrity, authenticity, and legal validity. PAdES is based on ETSI specifications and is adopted by the European Union and other countries, such as Brazil [(ETSI) 2015, ITI 2015].

2.2.1. Incremental Updates

It is possible to update a PDF file incrementally without rewriting the entire file. Changes should be appended to the end of the file, preserving the original content [ISO 2020]. The cross-reference section for an incremental update should only include entries for modified, replaced, or deleted objects. When reading the file, a PDF reader should prioritize accessing the most recent copy of each object from the file [ISO 2020]. In ISO 32000, each signature in a PDF is tied to a single signer and can only include one signing certificate. However, PDF allows for the inclusion of multiple signature dictionaries, each with its own `ByteRange` specification [(ETSI) 2015]. The `ByteRange` serves as a mechanism to distinguish between the signed and unsigned data blocks. In a typical serial signature, the document contains sequential signatures of different individuals. Subsequent signers not only add their own signatures but also sign the previously applied PDF signatures [(ETSI) 2015]. Figure 1 illustrates multiple signatures in a PDF file.

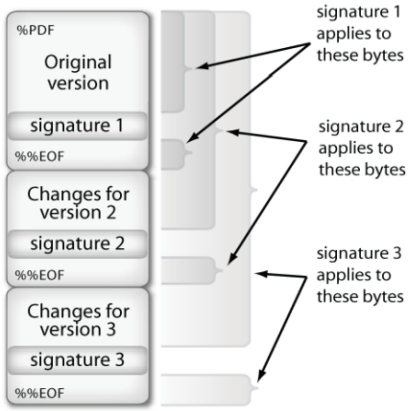

Figure 1. An example of serial signatures in the PDF structure:[(ETSI) 2015].

During the validation process, in compliance with the ISO 32000 standard [ISO 2008], each signature undergoes individual verification. The final determination of the document's status is made by aggregating the outcomes of all signature validations. It is essential to bear in mind that while certain signatures may successfully pass validation,

others might fail due to document alterations or concerns related to trustworthiness. As a result, the overall state of the document is ascertained by considering the combined results of the validation procedures [(ETSI) 2015].

2.2.2. Interactive Forms

An interactive form, commonly known as an `AcroForm`, constitutes a comprehensive set of fields intended for interactive data collection from users. Within a PDF document, a multitude of fields can be present across various pages, collectively forming a cohesive interactive form that spans the entire document [ISO 2020]. These interactive forms offer support for 'Text fields', which provide boxes or spaces where users can input text using a keyboard, as well as 'Signature fields', which represent digital signatures. The structure and attributes of a document's interactive form are defined by an interactive form dictionary, referenced through the `AcroForm` entry in the document catalog dictionary [ISO 2020].

Among these entries is the `SigFlags` property, which consists of flags that specify document-level characteristics associated with signature fields. Two particular flags fall under `SigFlags`: `SignaturesExist` and `AppendOnly`. The first, `SignaturesExist`, when enabled, allows the document to be modified by a form, while preserving the integrity of the signature. The second, `AppendOnly`, if set, signifies that if the form is saved (written) in a manner that modifies its previous content, any existing signature will be deemed invalid [ISO 2020].

2.2.3. Confidentiality in the General Data Protection Law

Law No. 13,709 [Brasil 2018], known as the General Data Protection Law, regulates the processing of personal data, including in digital environments, by natural persons or legal entities, whether public or private, with the aim of protecting fundamental rights of freedom and privacy. For the context of this work, Articles 6 and 46 stand out.

Article 6 outlines the principles governing the processing of personal data [Brasil 2018]. Among these principles, the following are particularly relevant to the context of this work: **necessity**: limitation of the processing to what is necessary to fulfill the purposes informed to the data subject; **free access**: guaranteeing, to data subjects, free and easy consultation about the form and duration of the processing, as well as about the integrity of their personal data; **security**: using technical and administrative measures suitable to protect personal data from unauthorized access, accidental or unlawful situations of destruction, loss, alteration, communication, or any form of inappropriate or unlawful processing; **prevention**: adoption of measures to prevent the occurrence of damages due to the processing of personal data.

Article 46 describes the security responsibilities of the data controller [Brasil 2018]. Data controllers must adopt security measures, both technical and administrative, capable of protecting personal data from unauthorized access and accidental or unlawful situations of destruction, loss, alteration, communication, or any form of inadequate or unlawful processing. The same article reinforces the need to adopt security measures from the design phase of the product or service to its execution.

3. Related Work

Among the related works, we distinguish two different categories, those that seek to provide information obscurity to protect sensitive document information, the Sanitization, and works that seek confidentiality in external storage.

Wu et al. (2022) proposed a hierarchical medical record storage model for storing encrypted data, such as patients' medical images, to ensure record confidentiality in the cloud. Wu's proposal aims to alleviate the storage and management burden of local data. In this approach, the local server randomly generates a secret key and then employs a traditional encryption algorithm to encrypt the data before sending it to the cloud. When queried by a physician, the local server is triggered to retrieve the encrypted data [Wu et al. 2022]. No storage optimization is utilized, and the proposed security and privacy are closely tied to the "server" since it has access to the encrypted data.

Goyal et al. (2006) developed a cryptosystem called "Key-Policy Attribute-Based Encryption (KPABE)" for sharing encrypted data. In this cryptographic system, encrypted texts are labeled with attribute sets, and private keys are associated with access structures that control which encrypted texts a user can access for decryption [Goyal et al. 2006]. Hence, the access policy is closely linked to a user and file management structure.

Research on sanitization faces significant challenges in improving techniques for selecting information within a document that needs to be removed or altered. The effectiveness of such systems relies heavily on the analyzing algorithm [Friedlin and McDonald 2008]. Due to the difficulty of manually detecting and eliminating all possibilities, various algorithms have been employed in the literature to automatically sanitize documents, including ERASE [Chakaravarthy et al. 2008], N-Sanitization [Iwendi et al. 2020], PACO2DT [Lin et al. 2021], and Btop [Chakaravarthy et al. 2008].

It is worth noting that research on sanitization typically does not address scenarios where the information requiring sanitization is well-defined, such as in the case of civil documents. These documents adhere to standardized templates where personal data is inserted. Furthermore, current research does not consider situations where there is a need for recovery of the original document, as the document is usually permanently altered.

4. Proposed Model

This section introduces a secure storage framework designed for the Brazilian Civil Registry, that can be easily extended to other contexts. This framework effectively tackles essential aspects, including safeguarding official documents' integrity through a dual digital signature mechanism and optimized storage that maintains the confidentiality of citizen data. Our model integrates cryptographic primitives, such as symmetric encryption and digital signatures, with a novel data storage and management strategy.

Our model also aims to facilitate the retrieval of digital documents by citizens, lowering the need for technical expertise or possession of cryptographic tokens. An inclusive and secure system should not impose the burden of comprehending intricate security principles on users seeking access to their digital documents. Additionally, it is imperative that the document remains accessible online, safeguarding the confidentiality of personal data contained within, and upholding the integrity of the official document.

First, we outline our proposed technique for issuing official documents and keeping them safe for eventual retrieval. Second, we present the procedure for obtaining documents from secure storage. Thirdly, a formal description is provided. Finally, this section is concluded with a discussion of security and regulatory matters.

4.1. Document Issuance

The proposed technique for issuing and storing documents by civil registrars consists of three components or systems: (i) registrar's office, (ii) secure platform, and (iii) secure storage. Figure 2 depicts these components and their relationships.

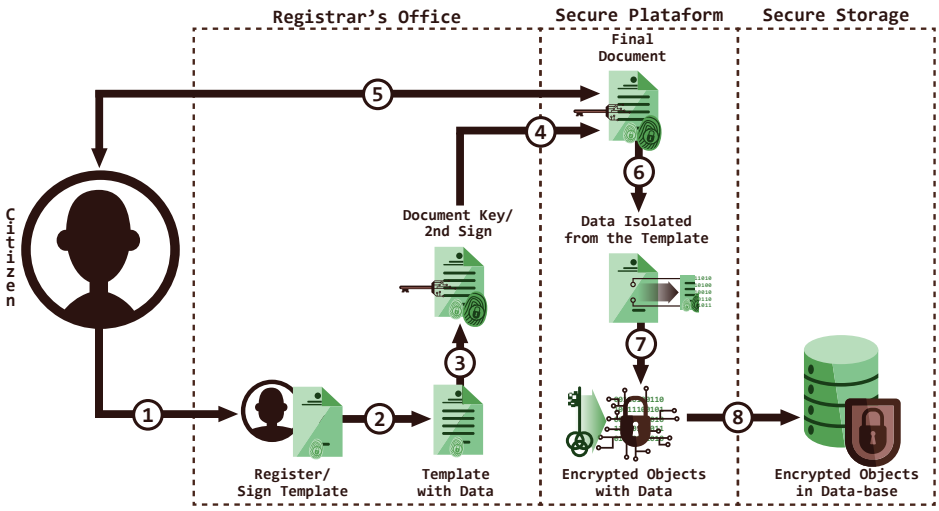

Figure 2. The process of safely issuing, signing, and storing a document.

The proposed model works as follows. In step 1, a citizen submits a request to a civil registrar to issue a civil document. After verifying the authenticity of the request and obtaining the data, both of which are beyond the scope of this work, the civil registrar inputs the citizen's personal information into an unfilled but signed document called a template in step 2. Then, in step 3, the civil registrar's system generates a pseudo-random number known as a document key, which is added to the document in a visible manner. Also, the civil registrar digitally signs the filled-out template in step 3, producing the requested official document. In step 4, the registrar's system transmits the document to the security platform for long-term archival and sharing with the user. In 5, the newly issued official document remains available to customers for a limited period. Both personal information described in the document and the registrar's signature are extracted from the document in step 6. The extracted data is encrypted using the secure platform key and a symmetric encryption scheme such as AES in step 7 and stored in a database in step 8.

In this system, three security assumptions are made. First, the template-filling procedure is carried out by a trustworthy civil registrar whose system has not been compromised. Second, the secure platform has access to a key for symmetric encryption and decryption and is trustworthy. Third, communication between the secure platform and the registrar's system is conducted via a secure channel. We claim that these assumptions are realistic and reflective of real-world systems. It is important to note that secure storage is

considered secure not because the database management system encrypts its content but because all of the data it stores is encrypted beforehand.

4.1.1. Dual Digital Signature

Our proposed method for ensuring the authenticity of official documents in the digital domain exceeds the existing literature. The approach centers on using two digital signatures to resolve the LGPD-mandated obligations of the data controller regarding data integrity (see 2.2.3). Additionally, the proposal improves the legal standing of said documents. We propose to use a dual digital signature scheme, which involves applying two digital signatures to the overall document: an initial signature performed by the entity responsible for issuing the official document template and a second signature, made after the inclusion of the citizen information, performed by a registrar with public trust.

Thus, we propose the creation of a PDF document with interactive forms (see 2.2.2), signed with a form-filling policy. The `SignatureExist` flag in `SigFlags` is used to accomplish this, ensuring that the modification of any form fields is not regarded as an incremental change. Furthermore, the registrar's signature sets the "no changes" policy of its signature, blocking any future modifications to the form fields and disallowing any additional digital signatures after document completion. This measure assures the security and immutability of citizen information. With these methods combined, we offer an efficient solution for creating, signing, and validating documents in Brazil's Civil Registry while maintaining the integrity of signatures via PDF signing policies.

4.1.2. Optimized Storage

Due to the vast number of documents issued by Brazil's civil registrar, it is essential to optimize storage space. According to official estimates, there is approximately 235 million registries at the moment of writing, with approximately 5 million new registries added annually [1]. To optimize the storage of issued documents, we propose retaining only the filled-in data and digital signature while discarding the template and its signature.

The proposed separation of filled data and the registrar's signature from the template is achievable because the PDF standard specifies that filled form data and signatures are appended to the document, *i.e.*, the template in our system. In other words, the final document delivered to the user results from appending the unfilled signed template with new objects and structures that represent the form data and the registrar's signature. As the template is standardized, it is possible to conserve storage space by discarding the template bytes and storing only the form data and the registrar's signature.

The procedure for separating the final document from the blank template depicted in step 6 of Figure 2 is as follows. Read the dual signed birth certificate from the n byte to the end of the file, where n is the unfilled template's size in bytes. These are the filled data and the registrar's signature. Additionally, Secure storage is achieved with a single-table database with two columns: index and extracted data. We describe the encryption of the extracted data to ensure confidentiality above (see 4.1.1). Now, let us explain how we

[1] `https://transparencia.registrocivil.org.br/registros`

maintain confidentiality for the index. The index comprises the document key (introduced in step 3 of Figure 2) encrypted with the secure platform key.

4.2. Document Recovery

After requesting the issuance of an official document, users may retrieve the document using the secure platform's document recovery process shown in Figure 3.

Step 1 involves entering the document key into an online interface. Step 2 encrypts the document key with the secure platform key to search the database with the index. Step 3 returns encrypted form data with the registrar's signature if a match is detected. Step 4 decrypts using the secure platform key to obtain the form fields and the registrar's signature. The dual signed birth certificate document is created in step 5 by concatenating the blank template with the deciphered data. In step 6, the user receives the document.

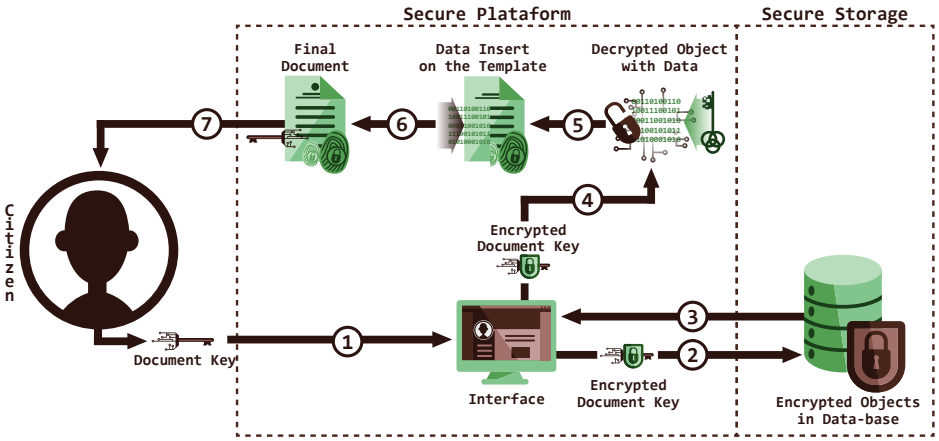

Figure 3. The process of recovering a document.

4.3. Formal Description

In this section, we will formally describe the algorithms that constitute our proposal. For brevity, instead of defining commonly used cryptographic primitives, let us discuss our notation. We employ KGen, a key generation function that receives a security parameter; HMAC_k is a Hash-based Message Authentication Code (HMAC) that uses key k; Enc_k is the symmetric encryption function that uses key k; Dec_k is the respective decryption function that takes key k; and Sign_k is a digital signature function that uses private key k. We define the following functions and symbols:

- $\mathsf{ek} \leftarrow\!\!\$\ \mathsf{KGen}(1^n)$ is the secure platform's symmetric key for encryption and decryption with security parameter 1^n;
- $\mathsf{dk} \leftarrow\!\!\$\ \mathsf{KGen}(1^{n'})$ is the document key with security parameter $1^{n'}$;
- sk: the registrar's private key;
- $template$: unfilled and signed template for a birth certificate;
- δ: citizen's personal data to be added to a birth certificate;
- $birth\_cert$: a dual signed birth certificate;

- $frac$: an encrypted fraction of a dual signed birth certificate.

Algorithm 1 describes the creation of the birth certificate by the registrar. On line 1, the dk is sampled. On the second line, the unfilled signed $template$ of the birth certificate is filled in with data from the citizen, represented by δ, and dk. On line 3, the now filled birth certificate is signed with the registrar's private key sk, producing the dual signed $birth\_cert$. These are steps 2 and 3 shown in Figure 2.

Algorithm 1: Dual signed birth certificate generation

Input : Empty $template$, person's private information δ, and the registrar's sk_r
Output : Digitally signed birth certificate $birth\_cert$

1 $\mathsf{dk} \leftarrow\$ \mathsf{KGen}(1^{n'})$;
2 $template' \leftarrow (template, \delta, \mathsf{dk})$;
3 $birth\_cert \leftarrow \mathsf{Sign}_{\mathsf{sk}}(template')$
4 **return** $birth\_cert$;

Algorithm 2 describes the process performed by the secure platform to extract, encrypt and store a fraction of the birth certificate in the database. On line 1, the registrar's signature and personal data are extracted from the dual signed birth certificate, producing $frac$. On line 2 we create an encrypted unique index to store and later retrieve $frac$ from the database using HMAC, which is named id. Refer to steps 6 and 7 shown in Figure 2.

Algorithm 2: Encryption of fractional document and index

Input : Empty $template$, signed $birth\_cert$, its dk, and the platform's ek.
Output : Encrypted data and signature $frac$ and encrypted database index id

1 $frac \leftarrow \mathsf{Enc}_{\mathsf{ek}}(birth\_cert - template)$;
2 $id \leftarrow \mathsf{HMAC}_{\mathsf{ek}}(\mathsf{dk})$;
3 **return** $frac$, id;

Algorithm 3 describes the document recovery process based on the dk entered by a citizen on the secure platform. Just like in Algorithm 2, dk is submitted to HMAC to create id on line 1. Then, line 2 verifies if the id exists in the database, if it exists, on line 3, the fraction of the document stored in the database is decrypted. On line 4 the birth certificate is fully recreated by concatenating the $template$ and the deciphered fraction. Refer to steps described in the recovery model, in Section 4.2 and Figure 3.

4.4. Discussion

The LGPD mandates that data controllers implement proper security measures to safeguard personal data from unauthorized access, alteration, destruction, or misuse. Our proposal ensures protection against alteration through digital signatures. We propose encrypting personal data in civil documents using symmetric encryption methods like AES (refer to step 7 in Figure 2). The encryption process utilizes the security platform's encryption key, which can be securely stored and operated in a Hardware Security Module (HSM). While our encryption mechanism guarantees security against unauthorized access and misuse, our proposal does not cover protection against destruction. It can be achieved through immutable storage, such as linking encrypted records using a hash function or HMAC.

Algorithm 3: Recreation of dual signed birth certificate $birth\_cert$.

Input : User-provided dk, the secure platform's ek, and unfilled signed $template$.
Output : A dual signed birth certificate $birth\_cert$ or $\emptyset$.

$id \leftarrow \mathsf{HMAC}_{\mathsf{ek}}(\mathsf{dk})$;
if *id exists in the database* **then**
    $frac \leftarrow$ retrieve $frac$ from the database using index id;
    $birth\_cert \leftarrow template \| \mathsf{Dec}_{\mathsf{ek}}(frac)$;
    return $birth\_cert$;
else
    return $\emptyset$;

It is important to note that the document key is printed on the document and exclusively accessible to the citizen, promoting minimal hardware and data privacy. The document owner can use the document key for various purposes, such as keeping a physical backup in case of loss or sharing it with someone else to have them carry out the recovery process, eliminating the need to handle digital or printed versions altogether. Nevertheless, requiring users to manage a lengthy document key may be perceived as a usability drawback. To address this, the document key size can be reduced and combined with other information, such as the national ID identifier.

5. Security Analysis

Storing sensitive citizen data requires careful security considerations. In this section, we quantify the probability of an adversary obtaining a birth certificate over some time, discuss the security properties introduced by our protocol, and analyze three attack vectors: (i) retrieval of a random birth certificate through brute force; (ii) retrieval of a specific birth certificate through brute force; and (iii) an adversary with access to a database dump. Furthermore, we present a set of security parameters with different trade-offs between usability and security. Let us begin by defining the security parameter θ.

Definition 1. *Let D be the number of documents, K be the number of possible $document\_keys$, α be the number of distinct password guessing per hour, then the security parameter θ is:*

$$\theta = D \times \frac{24 \times 30 \times \alpha}{K} \tag{1}$$

Preventing repeated authentication attempts by fraudulent actors without negatively impacting end users is a typical difficulty for online services. Each platform employs different methods to prevent unauthorized access to sensitive information. Standard methods include the completion of captchas, restricting the maximum number of authentication attempts before lock-out, and limiting the number of queries made. Each method has usability and security trade-offs that must be carefully considered when protecting a system. Our protocol limits the number of queries as it is the most adequate for widespread usability.

Theorem 5.1. *Let β be the maximum number of queries made per hour by a user by a platform P. Thus, there exists β such that an adversary cannot easily retrieve any specific information from P through brute force.*

Proof. From Eq. 1, we note the existence of parameters K and α, which directly influence the security level of platform P. Therefore, for a security requirement β and a particular D, there are a combination of parameters K and α that satisfies $\beta < \theta$. □

We highlight from Theorem 5.1 that brute force attacks can be made ineffective. Additionally, security is enhanced when searching for a specific birth certificate as the number of documents D is reduced to one. Consequently, the probability of revealing a specific birth certificate is also lowered as θ is proportional to D. Thus, we can comfortably assume that for a secure set of parameters, personal information is kept private.

We note that there are still trade-offs between security and usability. As this proposal is focused on widespread adoption by the population, we must consider lower levels of technical expertise; therefore, longer passwords and extended query restriction times might hamper adaptability. Thus, every implementation must decide on the optimal set of parameters to maximize usability and security for its environment. Consequently, Table 1 presents a different set of security parameters, adjusting password length and query limit through K and α correspondingly for three scenarios.

Table 1. Parameter recommendations for implementations focusing on different levels of usability and security with K document keys, assuming α number of distinct password guessing per hour, and security level θ for an environment with 203 million documents.

Usability	Security	K	α	θ
Low	High	36^{14}	60	1.42×10^{-9}
Mid	Mid	36^{12}	60	1.85×10^{-6}
High	Low	36^{10}	10	3.99×10^{-4}

Finally, we must discuss database leaks. We recall that the database is ciphered by a symmetric encryption key (ek); thus, any attacker possessing the encrypted database would require either the symmetric key or to break the underlying encryption algorithm. Therefore, we recommend a symmetric encryption method with at least 256 bits of security. In our implementations and tests, we opt for the AES-256 algorithm.

6. Implementation and Evaluation

In this section, we detail how our proposal was implemented and present empirical results.

6.1. Simulation Setup

With regards to the variables of the formal description, we used $n = 256$, which is the key length for AES-256-CBC, $h = 256$ for an HMAC based on SHA-256, and $n' = log_2(36^{12})$, which is a 12-character combination of numbers and lower-case letters.

To evaluate the proposed model empirically, we created a simulation using the micro-service architecture with Docker. The systems mentioned in Section 4 were developed as services that communicate through a TCP/IP network. The algorithms mentioned earlier were implemented in Python 3, and PostgreSQL 14 was used for the secure storage.

For our evaluation, we obtained an empty birth certificate document from Brazil's Civil Registry, which did not have any form fields. We created the fields and digitally signed

the template, resulting in a signed template of 450 424 bytes. Subsequently, we conducted an experiment where random names, dates, and other information were generated using the Faker library[2]. Our code is open source and available online[3].

6.2. Evaluation Results

The average size of a birth certificate, including the registrar's digital signature, was 472 684 bytes. The size of the filled data with the signature was 22 260 bytes, which increased by an additional 300 bytes after AES 256 encryption due to the inclusion of encryption-related information such as key size and AES's initiation vector. Thus, our proposal persists approximately 0.047% of the birth certificate. It is worth noting that compression algorithms can be applied before encryption to achieve higher gains. For example, using LZMA compression, the average amount of stored information can be reduced to 9010 bytes, corresponding to 0.019% of the birth certificate. With regards to compressing the entire birth certificate with LZMA, our proposal achieves a rate of 0.021% relative to the size of the compressed birth certificate, which is 423 559 bytes on average.

Our empirical experiment generated realistic birth certificates and compared the storage space required to persist them using our proposal versus compressing and storing documents individually. Figure 4 illustrates the total size of stored documents in the secure storage on the Y axis, plotted against the number of birth certificates on the X axis. Both axes are displayed in logarithmic scale. The top black line represents the storage size required when storing birth certificates without employing any technique other than LZMA compression (referred to as the naive technique). The purple bottom line represents our proposal, which requires almost two orders of magnitude less storage than the naive technique for a given number of birth certificates.

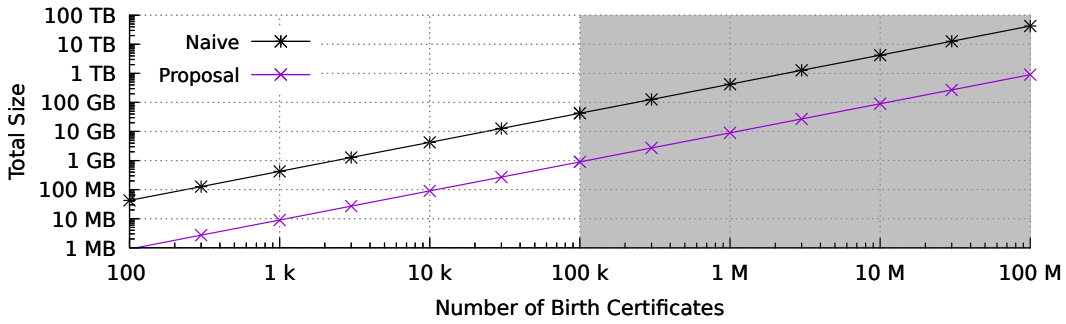

Figure 4. Storage requirements for birth certificates with logarithmic axes.

We empirically evaluated up to 100 000 birth certificates. The data points within the grey rectangle on the right side of the graph were estimated based on the empirically measured values of the left side. Using this estimate and assuming, for convenience, that the whole Brazilian population (203 million individuals) was born in Brazil and has a birth certificate, the naive technique would require around 78.20 terabytes to store all the certificates. Our proposed solution reduces this storage requirement to just 1.66 terabytes.

[2] https://github.com/joke2k/faker
[3] https://mega.nz/folder/yoM0GaDS#rPyFSKSzY0IqbX7vxARzww

7. Final Remarks

We present a new secure cryptographic protocol for the confidentiality of electronic documents, developed in accordance with ISO 32000 and PAdES standards and based on actual requirements for civil registration. Our protocol addresses both computational performance considerations and accessibility for citizens through the design of the protocol, ensuring that *free access* is maintained for any and all citizens. Furthermore, we demonstrate how the document key can be stored in a simplified way and show that it can be used in conjunction with a platform key to ensure document confidentiality according to the principles of *safety* and *prevention*. Moreover, we store a minimal amount of encrypted data according to the *principle of necessity*. Additionally, our proposal reduces storage requirements by a substantial amount. Notably, our findings indicate that only 1.66 TB of storage space would suffice to store all current Brazilian residents' birth certificates.

Future works. While this proposal concentrates primarily on the Brazilian civil registry case and the storage of citizen birth certificates, the model is flexible enough to be modeled after other use case scenarios, each of which has its own set of characteristics that must be addressed. Additionally, in this proposal, we have chosen document keys that are simple to remember in order to improve usability. However, there are numerous other ways to generate keys with sufficient entropy, so future works may opt to alter key generation methods in order to increase the security of generated document keys while maintaining acceptable usability. Finally, we do not examine the possibility of rogue entities in this study. Thus, it is possible to conduct a comprehensive security analysis of the entities themselves in order to detect and deal with malicious attackers within the protocol itself.

References

[Banerjee et al. 2017] Banerjee, A., Hasan, M., Rahman, M. A., and Chapagain, R. (2017). Cloak: A stream cipher based encryption protocol for mobile cloud computing. *IEEE Access*, 5:17678–17691.

[Brasil 1973] Brasil (1973). Lei de registros públicos - lei n. 6,015. Diário Oficial da União. Available in: `https://www.planalto.gov.br/ccivil_03/leis/l6015compilada.htm`.

[Brasil 2018] Brasil (2018). Lei nº 13.079, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). *Diário Oficial da União*, 157(1):59–64.

[Chakaravarthy et al. 2008] Chakaravarthy, V. T., Gupta, H., Roy, P., and Mohania, M. K. (2008). Efficient techniques for document sanitization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 843–852.

[CNJ 2011] CNJ (2011). Provimento nº 15. Available in: `https://atos.cnj.jus.br/atos/detalhar/1300`.

[CNJ 2012] CNJ (2012). Recomendação/orientação nº 6/2012. Available in: `https://atos.cnj.jus.br/files//recomendacao/recomendacao_6_02072012_22102012214535.pdf`.

[CNJ 2017] CNJ (2017). Provimento nº 63. Available in: `https://atos.cnj.jus.br/atos/detalhar/2525`.

[(ETSI) 2015] (ETSI), E. T. S. I. (2015). Electronic signatures and infrastructures: Pades digital signatures; part 1: Building blocks and pades baseline signatures. Technical report.

[Feng et al. 2017] Feng, S., Wu, C., Zhang, Y., and Oliva, G. (2017). Wsn deployment and localization using a mobile agent. *Wireless Personal Communications*, 97:4921–4931.

[Friedlin and McDonald 2008] Friedlin, F. J. and McDonald, C. J. (2008). A Software Tool for Removing Patient Identifying Information from Clinical Documents. *Journal of the American Medical Informatics Association*, 15(5):601–610.

[Goyal et al. 2006] Goyal, V., Pandey, O., Sahai, A., and Waters, B. (2006). Attribute-based encryption for fine-grained access control of encrypted data. In *Proceedings of the 13th ACM conference on Computer and communications security*, pages 89–98.

[ISO 2008] ISO (2008). Iso 32000: Portable document format (pdf). International Standardization Organization.

[ISO 2020] ISO (2020). Iso 32000-2: Portable document format (pdf) — part 2. International Standardization Organization.

[ITI 2015] ITI (2015). Documento icp-15 versão 7.0.

[Iwendi et al. 2020] Iwendi, C., Moqurrab, S. A., Anjum, A., Khan, S., Mohan, S., and Srivastava, G. (2020). N-sanitization: A semantic privacy-preserving framework for unstructured medical datasets. *Computer Communications*, 161:160–171.

[Khan et al. 2016] Khan, F., Li, H., and Zhang, L. (2016). Owner specified excessive access control for attribute based encryption. *IEEE Access*, 4:8967–8976.

[Li et al. 2022] Li, T., Wang, H., He, D., and Yu, J. (2022). Synchronized provable data possession based on blockchain for digital twin. *IEEE Transactions on Information Forensics and Security*, 17:472–485.

[Lin et al. 2021] Lin, J. C.-W., Srivastava, G., Zhang, Y., Djenouri, Y., and Aloqaily, M. (2021). Privacy-preserving multiobjective sanitization model in 6g iot environments. *IEEE Internet of Things Journal*, 8(7):5340–5349.

[Liu et al. 2017] Liu, J., Yu, J., and Shen, S. (2017). Energy-efficient two-layer cooperative defense scheme to secure sensor-clouds. *IEEE Transactions on Information Forensics and Security*, 13(2):408–420.

[Mohanty et al. 2016] Mohanty, M., Asghar, M. R., and Russello, G. (2016). $2dcrypt$: Image scaling and cropping in encrypted domains. *IEEE Transactions on Information Forensics and Security*, 11(11):2542–2555.

[Raja and Ramakrishnan 2020] Raja, J. and Ramakrishnan, M. (2020). Confidentiality-preserving based on attribute encryption using auditable access during encrypted records in cloud location. *The Journal of Supercomputing*, 76:6026–6039.

[Wu et al. 2012] Wu, Z., Xu, G., Yu, Z., Yi, X., Chen, E., and Zhang, Y. (2012). Executing sql queries over encrypted character strings in the database-as-service model. *Knowledge-Based Systems*, 35:332–348.

[Wu et al. 2022] Wu, Z., Xuan, S., Xie, J., Lin, C., and Lu, C. (2022). How to ensure the confidentiality of electronic medical records on the cloud: A technical perspective. *Computers in biology and medicine*, 147:105726.