Biometric Data for Archival Authentication

Alexander Aghili
Center for Research on Storage Systems
University of California, Santa Cruz
Santa Cruz, United States of America
awaghili@ucsc.edu
June 13, 2023

***Abstract*—Archival data must be stored and retrievable for decades to centuries. Protecting private data requires authentication methods that can last a long time and even potentially outlast the user. In this paper, we will examine utilizing biometric data from users to provide secure authentication capabilities for archives and provide a framework for implementing biometric authentication into an archive system.**

***Index Terms*—biometric, archive, authentication, biohashing**

I. Introduction

Archival systems store data for extended periods of time. Current systems of secure storage are unlikely to succeed over long timescales as password based RSA encryption comes with a host of issues. To secure data long term, an authentication method is preferably hard to steal and cannot be forgotten or changed. Biometric authentication fit these requirements better than any other current system of authentication. In this paper, we will examine how to design effective biometric authentication for archives and how such a system could be implemented. The paper is structured as follows: Section II identifies key issues for authentication issues in archival systems, section III discusses differing types of biometric data and storage techniques for each, section IV provides a framework for implementing biometric authentication into an archival system, section V summarizes additional challenges and complexities, section VI is the conclusion.

II. Authentication Issues in Archival Systems

Authentication in archival systems is more complex than in traditional systems. In an archive, there are long term consequences that a system designer with a shorter time horizon would not consider. To begin, hashing algorithms such as SHA-256 and cryptographic algorithms such as RSA are only computationally secure, meaning that with enough computational power and time, such encryption can be broken. Indeed, Peter Shor released a paper describing a method that would effectively break RSA encryption using quantum computers [1]. Additionally, issuing keys or indexes that must be stored by the user is often a poor choice in archival scenarios as those keys must be kept for a duration of decades or longer. Often, such keys are forgotten. Many people forget their passwords and have to be sent an email to set up a new password. This introduces a new problem, however, since as decades continue, email and other identifying accounts are lost and abandoned. In an archive, we seek to minimize the engagement required by the user, and differing recovery systems of variable effectiveness would not adhere to this requirement. These recovery services can also be maliciously accessed further degrading the stability of such key information. In the context of other authentication methods, such as biometrics, a system has to be able to handle the loss of certain information necessary to verify a user. For example, if someone used fingerprints to authenticate their data and then lost their hand, the system must still be able to recover their data. Even when the owner of the data has died, the data must still be recoverable.

These are important to consider carefully since archive data differs from most other data in that it must be highly secure to remain private. The lifetime of the data is often beyond the life of a single person, and as such, malicious actors have lots of time to plan and coordinate attacks. Additionally, if there is a leak or breach of data, or a broken authentication system, that exposes all the information thereby making one data breach of greater severity as compared to most other types of data.

Archival authentication isn't only required for individuals storing private data. Corporate and government data must be stored securely. Therefore, easy transitioning and revocation of authentication should be achieved for such entities in instances of both willing and forced alterations to the authenticated. One scenario might be an executive who is fired from a company, and wanting to prevent the leak of information, the archive must be able to revoke that executive's access to private data. Here, the owning legal entity of the data is the corporation. Another example is a corporation who goes bankrupt and its private intellectual property is sold to other entities. Here, the owning legal entity of the data shall be transferred to new owners. Archival systems must maintain an ability to comply with legal requirements regardless of the accessibility to the owner of such documents. This includes, but is not limited to, if there are legal government warrants for such data.

It is clear that providing authentication services for archival systems is no easy task and that careful design and implementation of any system is required to ensure the security of the data stored.

III. Storage of Biometric Data

Biometric data is a powerful tool to confidently identify a user without requiring storage or memorization of a key. Biometrics can range from hundreds of bytes to kilobytes making them very hard to guess. However, storing the raw biometric data is extremely risky as if a biometric is exposed, it is no longer a viable method of authentication. Biometric data is permanent to a user and cannot be reissued. Thus, it is essential that no data about the initial biometric can be leaked from any data stored to verify a user. Therefore, we shall utilize "cancelable biometrics" which performs transformations to biometric data to create a result which, even if exposed, does not reveal information about the biometric itself. An advantage of biometrics is that there is a variety of sources which will be discussed in this section. This means that even if one is compromised or otherwise lost, the others can still be used to provide authentication services. We shall now examine the differing types of biometrics.

A. Fingerprint

The fingerprint is a common biometric used in many industries to identify users and is generally unique across the population. The fingerprint provides 10 possible options for a single user (each finger) and is relatively easy access to scan with widely available technology to do so. However, loss of all fingers will render this biometric useless. Jin et al proposed a method of BioHashing to store a key for a fingerprint that accounts for noise [2]. It utilizes a pseudo-random token and many transformations, such as a wavelet transform and fast Fourier transform, that generates a feature vector. This feature vector can then be stored and the original biometric cannot be reproduced. However, having to keep and store a pseudo-random number to perform the transformations properly doesn't align with the goal of reducing the long-term storage requirements of the user. Syarif et al proposed a method of biohashing fingerprints that utilizes the most intensive histogram block location instead of a pseudo-random token making it less exposed to attacks and doesn't require the storage of any token for an extended period of time [3]. This improved method allows for fingerprint storage for archival purposes.

B. Eyes: Iris/Retina

The eyes have many unique characteristics that allow for identification. Iris and Retina scans utilize their uniqueness among a large subset of the populations. It is very difficult to steal these biometrics currently as scans are only successful within a close distance. This makes the Iris/Retina scan an ideal candidate for biometric authentication. The Department of Homeland Security, in conjunction with Clear Secure Inc, uses iris scans for identification at many airports across the U.S. Zuo et al published a paper detailing the methodology to create a cancelable biometric for iris scans [4].

C. Facial Recognition

Facial recognition technology has advanced greatly in the past few decades. Our entire personal lives, such as banking activity, personal messages, and social apps are hidden behind a facial identification scheme. The Apple iPhone uses facial recognition as authentication for the device and exposes all the earlier mentioned services, utilizing a 3D LIDAR scan of the face and storing that information on the device. Since the data near leaves the local device, there is no need for cancelable data storage. However, this method does conflict with archival requirements as phones get lost or broken. Ashiba et al provided a method of cancelable facial recognition [5]. Many countries and private companies maintain large databases of facial images for uses beyond pure authorization purposes.

D. DNA Analysis

DNA is a generally unique encoding of instructions in the cells that can be used to identify an individual. While there are potentially ways to extract out relationships between people, such as seen with modern paternity tests, from DNA, doing this requires storing the raw DNA data, thereby allowing for the identification of such relationships by the archive nearly impossible. Additionally, DNA can be identical between identical twins, it can be easily captured by a malicious actor from something as small as fallen hair, it can degrade over time, and it is currently expensive in terms of cost and time to process. Therefore, DNA is likely the worst biometric for archival use despite its identifying characteristics.

On top of the cancellation methods for specific biometrics, there are more general collection techniques for noisy data. One such method is *Fuzzy Extractors* presented by Dodis et al [6]. Fuzzy extractors can take a noisy input and turn it into a nearly uniform string with additional public helper data to generate a key. Li et al expanded on this, detailing a thorough end-to-end biometric identification and verification system [7]. Therefore, any of the methods used above can utilize fuzzy extractors to create cryptographic keys used in identification. There are alternative storage systems described by Tran et al to prevent exposing the original biometrics [8]. In the context of archives, though, general solutions come with the risk of exposing all of the biometric data if all encoding/encryption of the biometrics are done using the same process and it gets broken. It is more secure to utilize a different encoding scheme for each biometric.

Differing biometrics should be used in conjunction to simultaneously guarantee authentication in the event of a lost or exposed biometric and provide greater redundancy. This approach, referred to as multi-modal authentication, reduces both the False Acceptance Rate (FAR) and False Rejection Rate (FRR), key metrics which determine the error rate of the authentication system.

IV. Implementation of Biometric Authentication in an Archive

There are many challenges to implementing biometric authentication in archive systems.

A. Data Collection

Collecting biometric data can be a complex task. First, an archive must decide whether to require in person authentication or allow authentication over the internet. In person authentication does provide further assurance of the accuracy of the biometric and the authenticity of the user since all the testing and transmission can be done in isolated environment reducing the likelihood of successful attacks. This extra security comes at a great cost, however, as the archives must establish and maintain facilities that span the globe for reasonable access, hire people to staff those facilities, and verify proper procedures are executed among all facilities. Such a procedure would also effect the trust model of the archive since the system becomes centralized around the archives and their ability to provide authentication services. These extra costs makes it unlikely that an archive would set up these independent facilities, instead opting for utilizing biometric authentication that is already globally distributed.

The technology to collect all of the biometrics presented exists already. Devices such as the Apple iPhone or Microsoft Surface currently have facial recognition capabilities. Iris scanners are available at all the major airports across the United States. DNA collection is available in many medical settings using DNA sequencers. Previous iterations of the iPhone and every DMV across California have fingerprint scanners. Such devices are becoming more common as biometric authentication is utilized for general authentication services around the world. Therefore, the collection of the data can be done effectively on client devices. However, leaking the raw data risks exposing the biometric rendering it useless. Therefore, the cancelable transformations must be applied on a non-compromised client or the raw biometric can be sent to the authentication system over a secure channel utilizing encryption. The second option runs the risk of compromising the biometric if the encrypted data is collected and then the encryption method is broken.

B. Indexing

The implementation of the authentication system we propose will utilize the POTSHARDS archival storage system presented by Miller et al [9]. In POTSHARDS, data is stored across multiple archives using Shamir's Secret Sharing which splits the original data across archives in blocks, combined with additional random data, and has an index which traverses all the components to rebuild the original data. The client creates the index and is expected to store the index, using it to retrieve their data. In the case that the index is lost, POTSHARDS allows for approximate pointers, in which each piece of data has an indication of which large section of data the next piece will be in. This makes it such that retrieving the data via approximate pointers is time consuming and data intensive. Thus, a bad actor will be quickly detected while allowing for authorized access to the data in a specified time frame. However, this presents two problems. First, the archives require the user to store of an index. If the index is stolen, then the data can be maliciously accessed. In the event of a lost index, even with approximate pointers, the archive must ensure it only allows the collection of the data for authorized and authenticated users. Thus, we will utilize cancelable biometric authentication in a POTSHARDS system to allow for such capabilities.

In the event that a user doesn't want to store an index, then there is two main actions in the system. When the user wants to store the data and when they want to retrieve it. It is assumed that the communication channel is secure, meaning no information is leaked to a third-party. When the user wants to the store the data, there are three main steps. In the first step, the user will split the original data into chunks using Shamir's Secret Sharing (or have a service do it for them). Then, they will use biometric scanners to input their raw biometrics, which are quickly transformed into their cancelable versions using an open source algorithm. They will send one component along with the cancelable biometric to each archive. The archives will store the data and send back an index to the component. Step two involves the client creating a full index table from the indexes given by the archives and splitting the index into components using Shamir's Secret Sharing. That data is distributed among the archives and is sent with the cancelable biometric data. Step three occurs solely in the archive. Each archive, for all biometric data, should apply an additional unique but constant transformation to the cancelable biometric. This prevents identifying similar identities in the biometric data across archives. The cancelable biometric data stored in the archive points to a user identification key generated by the archive, and that user identification key points to index components. Thus, when an index component is added to an archive, the index is added in the list of components pointed to by the user identification key. All users will be able to access each starting index for all the data that user had access to in each archive. To prevent having to access all the stored data to figure out which index is the correct one the user is seeking, the connection from the user identification key to the index component can be named, creating a "title" for the data. Thus, the user only has to request that single title and the other indexes won't be rebuilt. Figure 1 shows an example of the first two steps of the process with three archives. Figure 2 shows the storage architecture of data in a single archive.

In retrieval, there are two core steps, which essentially reverse the storage processes. First, the user sends a request for an index component to each archives using their biometric, which has been transformed in the same manner as it was stored. The archive matches the biometric to a user

identification key, which then directs to a index component. That component is sent back to the user. When the user has all the components, it rebuilds the index. In step two, the user makes a request to each archive with the index they need and the biometric to authenticate their request. The archive verifies the user has authorization to access the index and then returns that component of the original data back to the user. Finally, upon all archives completing the requests, the user can rebuild the original data. Figure 3 provides an overview of the retrieval system with an example. As the components to rebuild the data and index are split among the archives in the same way, it would still require a significant amount of collusion between archives to access private data, thus making the proposed method no less secure than the original POTSHARDS proposal.

If the archive or user decides to utilize a user stored index, biometric authentication still provides significant benefits. The user identification key or the cancelable biometric can be stored as metadata within each component of data to ensure that the user attempting to access it has authorization to do so. This is especially useful in the case that a user has lost the index and needs to verify that they are who they say they are. This system would be similar to the previous without having to store the index. Figure 1 Step 1 and Figure 3 Step 2 is an example of such a system. The approximate pointers idea presented in POTSHARDS can still be used if a legal authority has a right to the data without requiring the presence of the user.

To ensure that the system can withstand the death of a user, multiple user identification keys can access the same index in the archive. Thus, an owning user can allow another user access to the data they stored so long as that other user has their biometrics inside of the system. To make giving access to others easier, the archive can store a table of all of the user's full legal names and other identifying information along with their cancelable biometrics such that an owner can easily query that user and add them to an access list. However, this exacerbates privacy issues and allows attacks to form connections between data using only metadata like user identification and cancelable biometrics associated with a specific user.

V. Further Complexities

There are additional considerations and complexities in building archival biometric authentication systems. Key assumptions made in the implementation stage were a secure communication channel and an open source transformation. A secure communication channel in which no data is leaked to a third-party between the archive and the user is necessary for this system to operate. If any data is leaked in the transaction, then the authentication system and data can be easily accessed. However, this concern is likely to be addressed by utilizing the state of the art secure communication methods of the day. The use of open source transformations to the biometric data is important because that same transformation must be applied for the user every time to be able to retrieve their data. If the transformations are closed source and the algorithm is out of date or forgotten as transformations are upgraded, the data protected by a biometric with the original transformation is lost.

A great concern among the general public is privacy. With cancelable biometrics, the data stored is designed to be irreversible and secure. However, it does not alter the notion that giving away our identities to be stored on a server is dangerous. This fear could inhibit enrollment in the archive system reducing its effectiveness. This is especially an issue if there is a way to extract the original biometric from the cancelable version. While the purpose of the cancelable biometric is to be irreversible, it could potentially be possible with enough computation and attempts, that certain cancelable biometrics are broken. This would render much of the protections associated with biometric data useless and further destroy the purpose of the secret sharing preformed in POTSHARDS. Thus, it becomes a single point of failure with respect to security and must be carefully implemented. Multi-modal identification mitigates this issue but the fact remains that preventing the exposure of original biometrics is paramount to a secure system.

VI. Conclusion

This article presented a biometric authentication system for archives, discussing differing biometric types, authentication designs, and storage techniques. In the context of archive systems, the key for an effective biometric authentication system is the guarantee that any biometric data stored is irreversible and cancelable. The methods of cancelling biometrics presented are designed to effectively utilize biometric data while preventing leakage of information that could expose the biometric. The implementation presented then builds upon the biometric data to fit into the archival context. For archival systems, piggy-backing off the technological development of other industries is more cost effective, a key component in archival systems. As biometric data is used in a greater capacity to identify individuals and biometric measurement systems become more widespread, it could be beneficial for archives to adopt using such technology to ensure the longevity of authentication. A biometric authentication system for archives would help to secure data for generations while allowing easy access to all the data one has a right to.

References

[1] Shor, P. W. (1997). Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer. SIAM Journal on Computing, 26(5), 1484–1509. https://doi.org/10.1137/s0097539795293172

[2] Jin, A. T., Ling, D. N., & Goh, A. (2004). Biohashing: Two factor authentication featuring fingerprint data and tokenised random number. Pattern Recognition, 37(11), 2245–2255. https://doi.org/10.1016/j.patcog.2004.04.011

[3] Syarif, M. A., Ong, T. S., Teoh, A. B., & Tee, C. (2014). Improved biohashing method based on most intensive histogram block location. Neural Information Processing, 644–652. https://doi.org/10.1007/978-3-319-12643-2_78

[4] Zuo, J., Ratha, N. K., & Connell, J. H. (2008). Cancelable iris biometric. 2008 19th International Conference on Pattern Recognition. https://doi.org/10.1109/icpr.2008.4761886
[5] Ashiba, M. I., Youness, H. A., & Ashiba, H. I. (2022). Suggested wavelet transform for cancelable face Recognition System. Multimedia Tools and Applications, 81(30), 43701–43726. https://doi.org/10.1007/s11042-022-13070-0
[6] Dodis, Y., Reyzin, L., & Smith, A. (2007). Fuzzy extractors. Security with Noisy Data, 79–99. https://doi.org/10.1007/978-1-84628-984-2_5
[7] Li, N., Guo, F., Mu, Y., Susilo, W., & Nepal, S. (2017). Fuzzy extractors for biometric identification. 2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). https://doi.org/10.1109/icdcs.2017.107
[8] Tran, Q. N., Turnbull, B. P., & Hu, J. (2021). Biometrics and privacy-preservation: How do they evolve? IEEE Open Journal of the Computer Society, 2, 179–191. https://doi.org/10.1109/ojcs.2021.3068385
[9] Storer, M. W., Greenan, K. M., Miller, E. L., & Voruganti, K. (2009). POTSHARDS—a secure, recoverable, long-term archival storage system. ACM Transactions on Storage, 5(2), 1–35. https://doi.org/10.1145/1534912.1534914

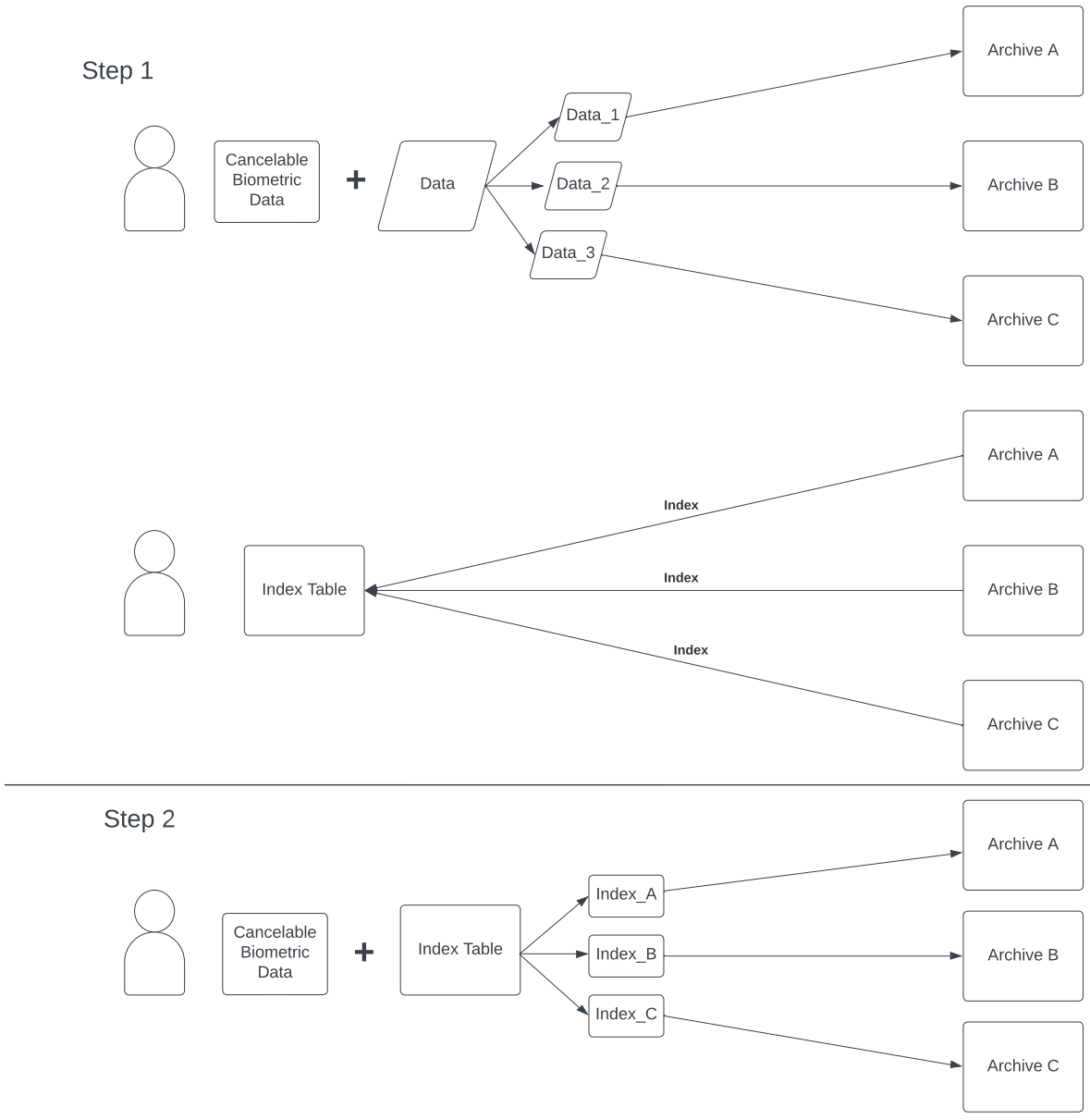

Fig. 1. Data Storage Process with Biometric Authentication

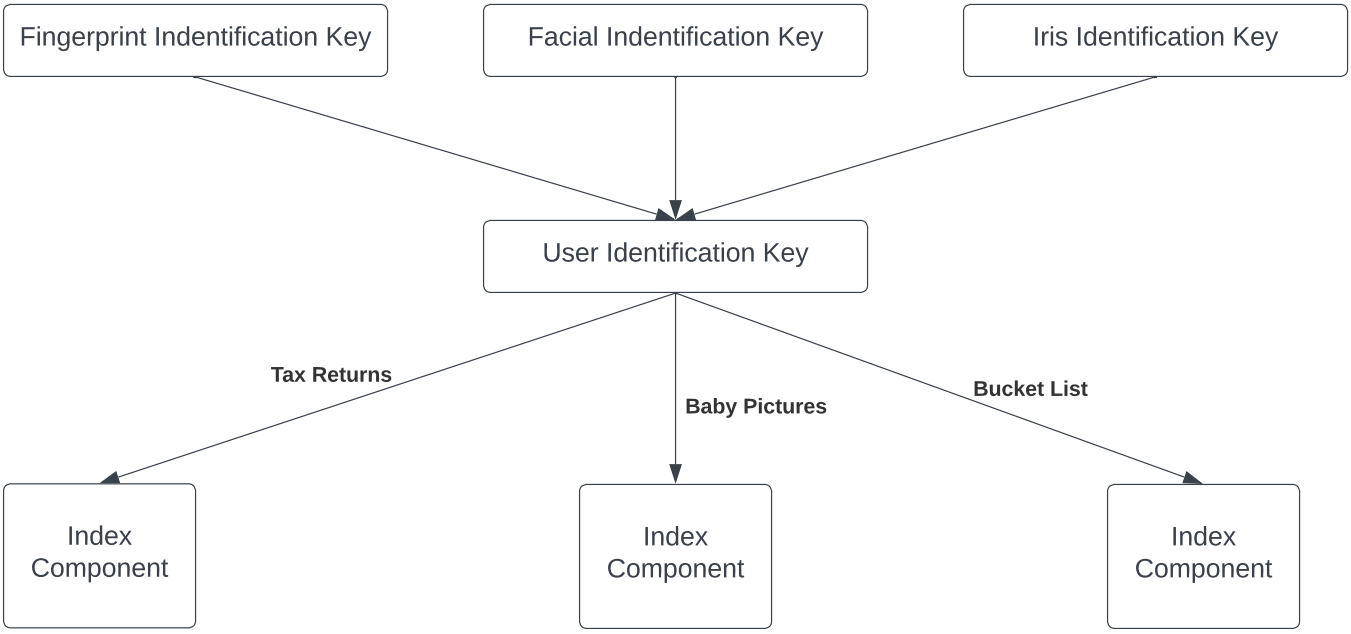

Fig. 2. Mapping Architecture

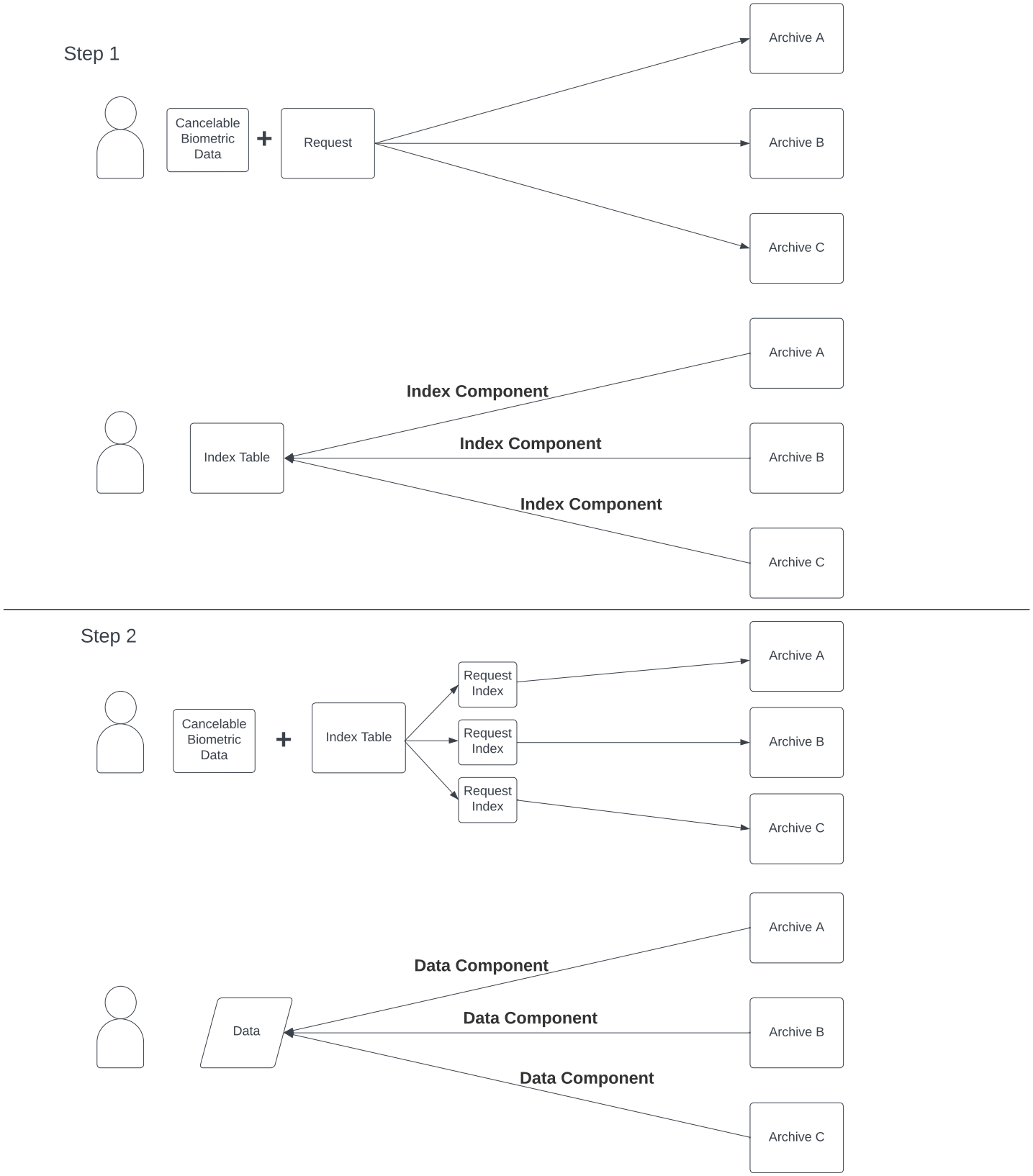

Fig. 3. Data Retrieval with Biometric Authentication