A

15 éves PEME XVII. PhD - Konferenciájának előadásai

(Budapest, 2018. november 15.)

Elektronikus könyv

BUDAPEST

2018

ISBN 978-615-5709-05-0

Kiadja a Professzorok az Európai Magyarországért Egyesület

TIZENÖT ÉVE AZ EURÓPAI SZINTŰ TUDOMÁNYOS MEGÚJULÁS ÉS A FIATAL KUTATÓK SZOLGÁLATÁBAN

2018. november 15.

BUDAPEST

Szerkesztette: Dr. Koncz István – Szova Ilona

Roskó Tibor/Adamkó Attila: The human DNA can be the bridge between the Human and its data set in the Future

Lektorálta: Dr. Gyarmati Péter és Dr. Mező Ferenc

Első szerző
Roskó Tibor
Debreceni Egyetem
rosko.tibor@inf.unideb.hu
ORCID: https://orcid.org/0000-0002-6521-9447

Második szerző
Adamkó Attila
Debreceni Egyetem
adamko.attila@inf.unideb.hu

Abstract

In our research, we are designing the abstract models of the global, centralized user authentication infrastructure. The user has only one trusted profile and a globally unique ID in the infrastructure. In this infrastructure model, we would like to support the cooperation among existing authentication providers. They can connect to the infrastructure if they comply with the common conditions.

To identify a person in a trusted way, we have to relate the physical person to its digital data with a trusted method. Nowadays, this trusted method is the facial image of the person. Unfortunately, the biometric properties are very damageable, so a general biometric property based databank is also very damageable. In this paper, we would like to describe our possible solution based on the Human DNA for this problem.

Keywords: FBI CODIS, DNA fingerprinting, DNA STR, integenX RapidHIT system, global centralized user authentication, national DNA-fingerprint databank

1. Introduction

In the 21st century, almost all the daily processes are managed via the Internet, online environment.[1] The biggest challenge in this is validating the identity of a user and compliance with the growing data privacy and other regulations on the side of both service providers and users. Nowadays, we have to register an account to all service before starting a user process. This method is a little bit old and not sustainable in a fast life as we are living, where everyone wants being served out immediately.[2] We are designing a global, centralized authentication infrastructure in our base research but it is a big theme with many ingredients. In this paper, we would like to give a possible solution for validating the identity of a physical user entity in the digital space. Unfortunately, it is not an easy method because there are many providers for authenticating a user but in common with them is using local, own registrations and mostly these are not based on a strict user identity validation. So, the job is building up a common way to check the identity of a user from a trusted source without using different methods if a chosen check option is not available for all. In other words, we have to find a common identity validation method which does not exclude groups of users by any reasons, for example, a user cannot give a fingerprint sample.

Our choice is the Human DNA (Deoxyribonucleic acid), exactly the DNA-fingerprint because it is the most commonly available for all users. Fortunately, it is not possible that a person cannot give a sample, in an extreme example, if we chop a user, even though we can identify it. In this paper, we would like to define the basic biometric user identification method and talk about the mostly used technics, for example, fingerprint, iris-retina, face and -naturally- DNA based methods. We would like to give an

abstract model about DNA-fingerprint based user identification compliance to the strict data privacy regulations with unique solutions. We would like to give some possible application contexts of the Human DNA from our viewpoint and define the limitations of the Human DNA applications.

2. The relation between the physical user and its digital data as built up today

The most important part of the user authentication is connecting the physical user to its digital data in a permanent, trusted way. If we are not able to reliably identify a user, we do not identify it just a vision about who is that. This cannot be the appropriate base for a trusted service, for example, a government or bank service but nowadays, we are sure it is not a passable way for simple service providers too.
Instead of visions, we are using a general method to build up the relation between the physical user and its data. This is the facial image. In this way, the facial image of the physical user connects the user to its basic information, for example, maiden name, date of born. This solution is suitable first for physical identification process maiden by human resource but widely used in the digital environment for validating the identity of the user too. For example, nowadays, we can register bank account online via video-checking method, it is based on a human resource matches our real-time face with our any identity card via video chat and naturally, special infrastructure which supports the process.[3]

Well, our physical entity is connected to our digital identity with our facial image. There are advantages of this method and unfortunately, disadvantages too.

The advantages of the facial image method in person identification:

- It is a widely used, well-specified option for connecting the physical entity to the digital entity.
- It does not need special hardware, just a simple camera, which is available in most mobile phones. If the matching process is managed by human resource, there is no needed special software.

The disadvantages of the facial image method in person identification:

- Unfortunately, as almost all biometric identifier of the human, it is very vulnerable. For example by fire or car accident damages.
- The facial image profile has to be updated in predefined time periods form the start of life until the end of life because the facial image profile continuously changing in different life stages.
- If the matching process is managed by computer resource, it needs special software.

As we mentioned, the facial image biometric identifier is mostly used for connecting the physical person to its digital data in the real world, mostly via identity documents, like passport, identity card. Thanks to the intensively developing software background, the officially captured and managed facial image profiles are also widely used for forensic identifications, for example in China. In first, in Hungary, there is also a regulation about creating facial image profiles based on the facial images of identity documents. This is the 2015. CLXXXVIII. regulation about the database and methods of the facial image analysis.[4] This system is based on the government personal data register databank, the up to date facial images of people are used to create up to date facial image profiles for mostly forensic recognitions, for example identifying a lost person, any unknown arrested person. The Chinese solution is the same way to use official registers for extended functions but it is a little bit more extended. They are using lots of security cameras (more than 170 million) to recognize people outside on the streets, the first goal is finding criminals. With this huge camera infrastructure, the Chinese police also hunt for general offenders and naturally, sanctions them, for example, if you go across the red lamp on foot, you will be recognized and almost surely sanctioned.[5][6][7]

3. The general definition of the biometric identifiers

"Biometrics is the technical term for body measurements and calculations. It refers to metrics related to human characteristics. Biometric identifiers are often categorized as physiological versus behavioral characteristics. Physiological characteristics are related to the shape of the body. Examples include, but are not limited to fingerprint, palm veins, face recognition, DNA, palm print, hand geometry, iris recognition, retina and odor/scent. Behavioral characteristics are related to the pattern of behavior of a person, including but not limited to typing rhythm, gait, and voice. Some researchers have coined the term behaviometrics to describe the latter class of biometrics."[8]

We have to separate the user recognition process two sharply different groups: identification and authentication. In our paper, we are working with the identification. When it is defined who is a user, we are talking about the identification. In this process, the user is described by its natural data: name, born information, mother's name, address, etc. If the question demonstrate who you are is answered by the user, we are talking about the authentication.[9] In this situation, the user has to demonstrate who is that, for example with a password, an identity card.

The biometric recognition solutions are also separated by availability and damage-ability metrics. Unfortunately, most of the methods are very damageable because these are the part of the outer body and they are exposed to be damaged in any accidents. For example fingerprint, facial image, vascular. The other problem with these options is that significant part of the people cannot give a sample for the identification process by various reasons. Well, these biometrics are widely used in identification and authentication processes but they are not absolutely stable and reliable.[10]

The almost only one possible solution is the DNA for meeting high availability and damage-proof conditions. The Human DNA is available for everyone, only one Human cell is enough that a person can give a sample. It is immune to chopping the sample person as we mentioned this extreme example of biometric sample damages.

Well, biometrics are very useful and trusted ways to identify a person but we also have to take care of many regulations. The most important is data privacy, the famous GDPR (General Data Protection Regulation). This is a common guide for managing people' personal data in the European Union but its definitions can be used worldwide. For biometrics, the most important section of the GDPR is that these biometric data are special personal information.[11] It means the user has to give explicit consent to the data processing. Unfortunately, the way to use biometrics for general identification is not the most obvious because it might be unconstitutional to mandatory store any of biometric properties. Against this, we would like to demonstrate the power of the almost best biometric, the DNA for officially identifying people, naturally, this solution has no goals to bring the DNA profile as a mandatory information in the personal register, such as name or birthdate.

3.1. Fingerprint in details for identifying people

The Human fingerprint based person identification is the most common way to identify someone in the physical and in the online environment. This method is used in most identity documents, such as passport, identity card so we can really say, it is a widely applied method for checking people's identity.[12] It is used to identify and authenticate a user in the digital environment but it is also used in forensic for identifying the criminals. But how does it work? The figure 1 below presents the simple process of the fingerprint recognition:[13]

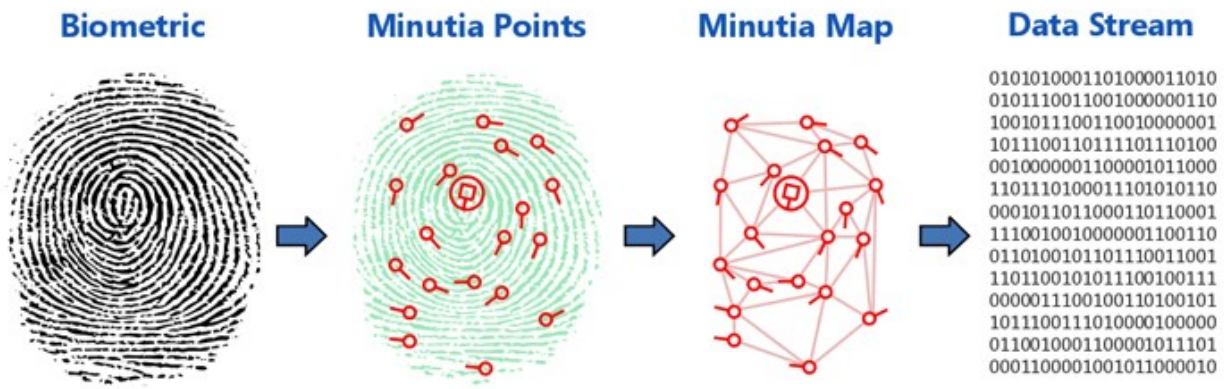

Figure 1 : The process of fingerprint recognition (source: http://www.identityone.net/BiometricTechnology.aspx)

The most important thing is that fingerprint image is not stored just a map about unique parameters of a fingerprint. This map will be used to match the given sample to the known fingerprint profile of the user.[14]

The advantages of the fingerprint method in person identification:

- It is almost persistent for a lifetime.
- From a certain point of view, the fingerprint can be easily read and checked, for example relatively fast matching for gate access control system.
- It is relatively unique for all people.

The disadvantages of the fingerprint method in person identification:

- It is very damageable.
- Not all people can give a sample by common reasons, for example, the person has no finger, the person's fingerprint unreadable.

3.1.1. Statistics about the fingerprint

We can say that the fingerprint is a unique biometric property of the Human but its uniqueness depends on the characteristics of the reader devices and recognition algorithms. There are four types of readers: optical, capacitive, ultrasound and thermal.[15]

The best available technology is the ultrasonic fingerprint reader because it uses high-frequency sound to recognize the fingerprint and by this property, it is very difficult to fool this sensor. It is more expensive than optical sensor but its properties bring that it worth its price.[16]

3.2. The most important measure is the FAR, FRR numbers to compare biometrics.

"The false acceptance rate, or FAR, is the measure of the likelihood that the biometric security system will incorrectly accept an access attempt by an unauthorized user. A system's FAR typically is stated as the ratio of the number of false acceptances divided by the number of identification attempts. (definition by Webopedia)"

"The false recognition rate, or FRR, is the measure of the likelihood that the biometric security system will incorrectly reject an access attempt by an authorized user. A system's FRR typically is stated as the ratio of the number of false recognitions divided by the number of identification attempts. (definition by Webopedia)"[17]

Along these measurements, we can classify the different biometric solutions, it is presented in figure 2 below:

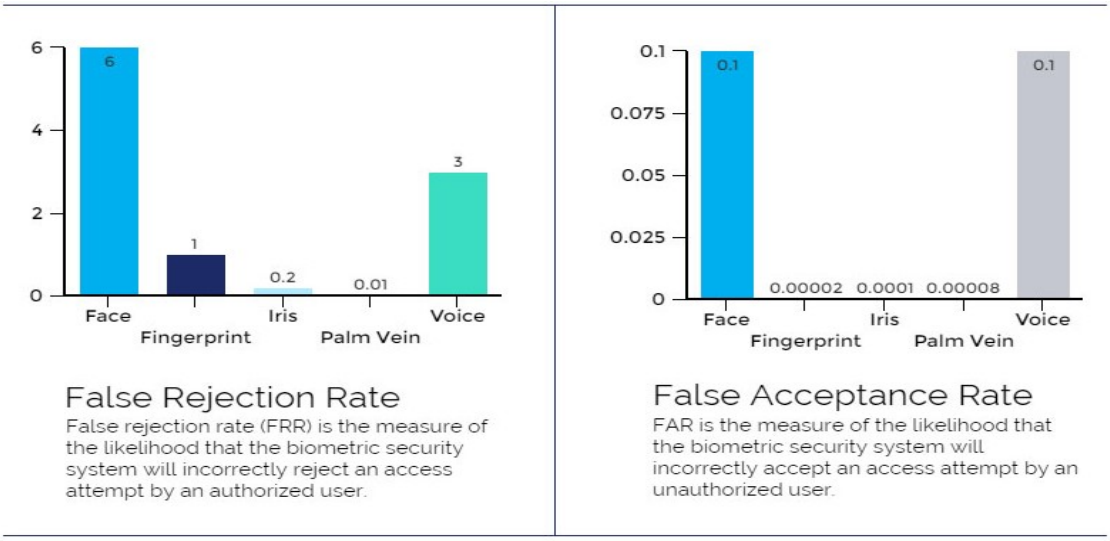

Figure 2 : The statistics of the FAR and the FRR (source: https://www.bayometric.com/biometrics-face-finger-iris-palm-voice/)

3.3. The description of the DNA, DNA-fingerprint

The previously described methods of biometrics are usually used to both identify and authenticate a user. But the next biometric method is a little bit different from them, this is none other than the DNA, exactly the DNA-fingerprint. It is widely used in forensics to identify criminals or lost people.[18] This solution is mostly for identifying instead of authenticating a person. In fact, the authentication process will be a little bit too long because a person identification process can be done more than 90 minutes with the best infrastructure too.[19]

The definition of the DNA: "deoxyribonucleic acid, is the hereditary material in humans and almost all other organisms. Nearly every cell in a person's body has the same DNA. Most DNA is located in the cell nucleus (where it is called nuclear DNA), but a small amount of DNA can also be found in the mitochondria (where it is called mitochondrial DNA or mtDNA). The information in DNA is stored as a code made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Human DNA consists of about 3 billion bases, and more than 99 percent of those bases are the same in all people. The order or sequence of these bases determines the information available for building and maintaining an organism, similar to the way in which letters of the alphabet appear in a certain order to form words and sentences."[20] These content is divided into coding and noncoding sequences. In our viewpoint, the important sequence is the noncoding sequence.

The widely used forensic testing method is the DNA-fingerprint, the next description presents how it is built up:

The most common investigating method is the STR (Short Tandem Repeat) which investigates predefined locis in the DNA chromosomes. Figure 3 below presents the process from the sample given until the profile outgoing:

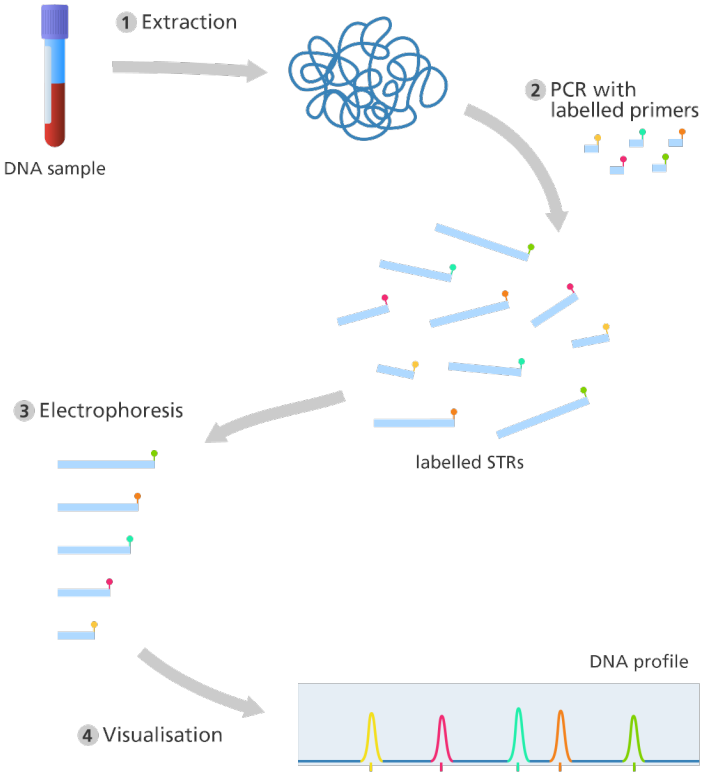

Figure 3 : The process of the DNA-fingerprinting (source: https://www.yourgenome.org/facts/what-is-a-dna-fingerprint)

Nowadays, the FBI (The Federal Bureau of Investigation) manages the CODIS (The Combined DNA Index System) for identifying criminals. The CODIS system became an international method for managing DNA-fingerprints of criminals, the commonly used loci packages are defined by the NIST (National Institute of Standards and Technology). It is a very important fact of the DNA-fingerprinting because the process is a well standardized, borderless applicable method for identifying people (not only criminals). Unfortunately, the base loci package of the CODIS became obsolete, so it was renewed in 2017 by adding new locis.[21][22]

The advantages of the DNA-fingerprinting in person identification:

- It is widely available for everyone and everyone is able to give a sample in extreme situations too, for example, if we chop the person, its DNA sample stays available.
- The DNA and the DNA-fingerprint are possibly unique for all.
- The DNA is one of the biometrics that can be applied not only for Human but for example animals too.

The disadvantages of the DNA-fingerprinting in person identification:

- The DNA is a special sample, to be investigated special devices are needed, and unfortunately, these devices are more expensive than other biometric systems.
- Nowadays, the DNA-fingerprint cannot identify the identical twins, they are not different based on their DNA-fingerprint.

4. Our plans to identify a person with the DNA-fingerprint

Our identification method research based on the STR based DNA-fingerprinting because it can provide a unique, widely available identification factor for connecting the physical person to its digital data. As it can be seen in the fig. 2, the nowadays used facial image identification method is not the best choice to identify a person. The FAR value represents that an algorithm might accept other people instead of the real user but the problem is the same when a physical person takes the identification process. The other problem with the facial image based identification is that the facial image model has to follow the changes of the facial image, so it has to be periodically updated.
In the next, we would like to describe our models for building up a global relation system between the Human and its digital data.

As we mentioned, the DNA-fingerprint is used mostly for identification, we also would like to use it to identify the users but not in forensics. Our idea is that build up national DNA-fingerprint databank which connects the physical person to its digital data. The basis is ready for this project but there are some questions and problems have to be solved. The most important is the data privacy adequacy. Our simple algorithm brings a solution that can take care about the privacy of the user and build up a comfortable system for fast identification in various cases.
Figure 4 below describes the model of our DNA-fingerprint databank management from the sample incoming until the DNA-fingerprint hash generating:

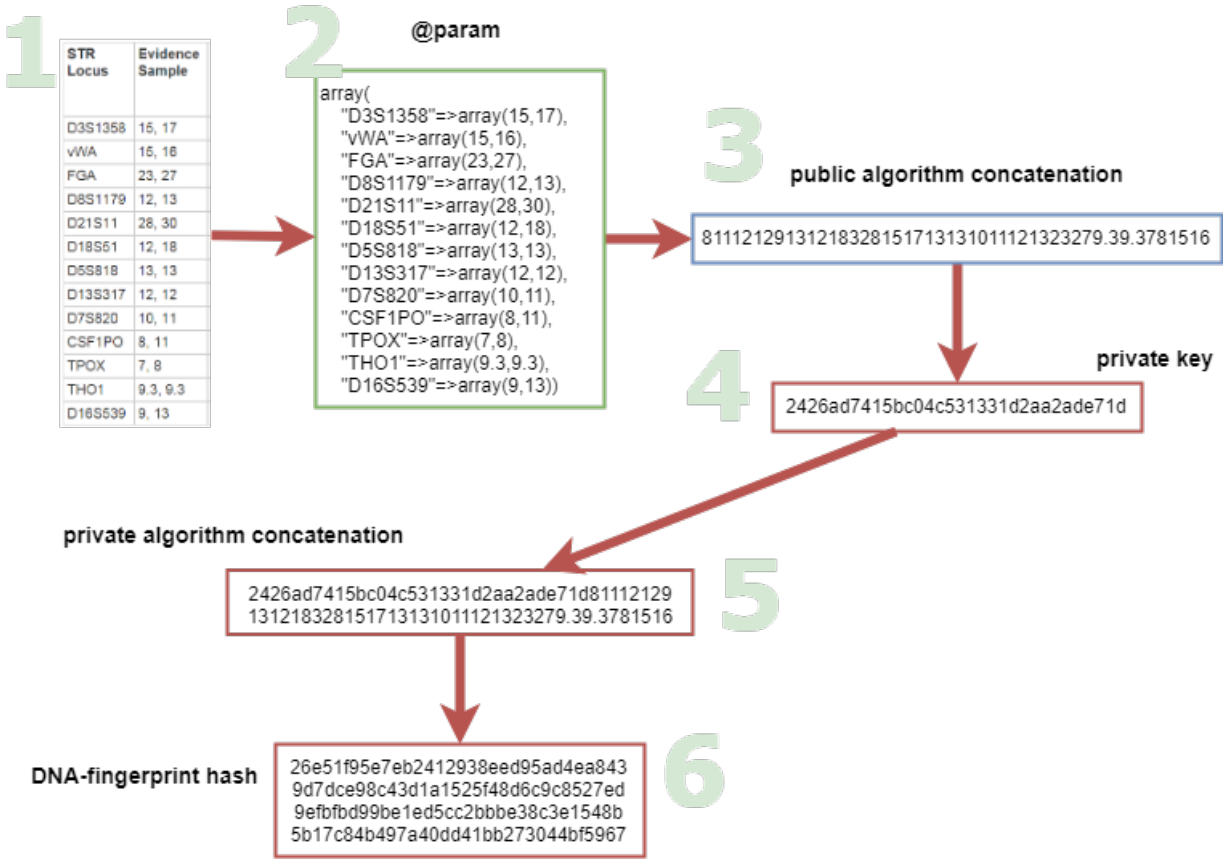

Figure 4 : The abstract model of our algorithm (source: own)

Immediately investigate the pseudos of the algorithms:

1. In first, predefined formatted parameters are given to the algorithm.

2. The next step is when the algorithm concatenates the given parameters, also into a predefined format. It is an important step because we want to repeat the hash generation in the identification process more times but if we do not use the same concatenation schema then the hashes will not be the same for the same sample.
3. And finally, the algorithm generates the hash for the previously concatenated values. The algorithm uses the same hash method, in this example we used the sha-3 hash algorithm with 512 length.[23] In the implementation of the national DNA-fingerprint databank, we recommend choosing the available best, robust hash method to avoid the known vulnerabilities, for example, collision attack.[24]

To raise the security of the hash generation, we can build up a little bit more complex system. In our idea, we can build up GDPR-compliance systems, if we are using private salt values for generating hashes. These private salt values are known only by the generator algorithm and stored in a "safe place".
These security steps are the public and private algorithms and the private keys. Figure 4 above is extended with these security options.

4.2. About the data privacy
As we mentioned the importance of data privacy, we also would like to give some solutions to comply with these regulations. The key of the solution is in the hash generation because it provides us a one-way encoding of the personal data. This cannot be decoded without the private keys which are known only by the algorithms. This complies to the pseudonymization option of the GDPR. It means the information cannot be known without special or more information, for example, private key. It also can be the base of the further usages when the fingerprint database will be a relation between other databases for extended usages, for example identifying an unknown person from its DNA-fingerprint (when the sample is fit for clear results).

4.3. The uniqueness of the DNA-fingerprint
The uniqueness is determined by the number of used locis in the DNA-fingerprinting. As we described, the forensic CODIS system uses 20 locis but the beta version contains 23+1 locis, it is possibly enough DNA-fingerprint can be unique for all except the identical twins.

The most important question is about the DNA-fingerprint uniqueness for identical twins. Our research plans are finding a suitable solution for this problem in a cooperation with DNA professionals. There are some new methods to also identify the identical twins but these are not common yet.[25]

4.4. The definition of the 4T data
The 1996. XX. regulation defines the basic attributes for uniquely identifying a person, these are:[26]
- actual last name and first name
- maiden last name and first name
- mother's maiden last name and first name
- place of birth
- date of birth

The subset of the 4T data is used in our research:
- maiden last name and first name
- mother's maiden last name and first name
- place of birth (ISO 3166-1 alpha-3 country code and the city name)[27]
- date of birth (format: YYYY-MM-DD)[28]

5. Further applications based on the national DNA-fingerprint databank

As we mentioned, the basic national DNA-fingerprint databank can be a great base for build up relations between different national databanks and prevent duplicates of one person's record. Along these relations we can develop various applications, the next, figure 5 describes a possible one:

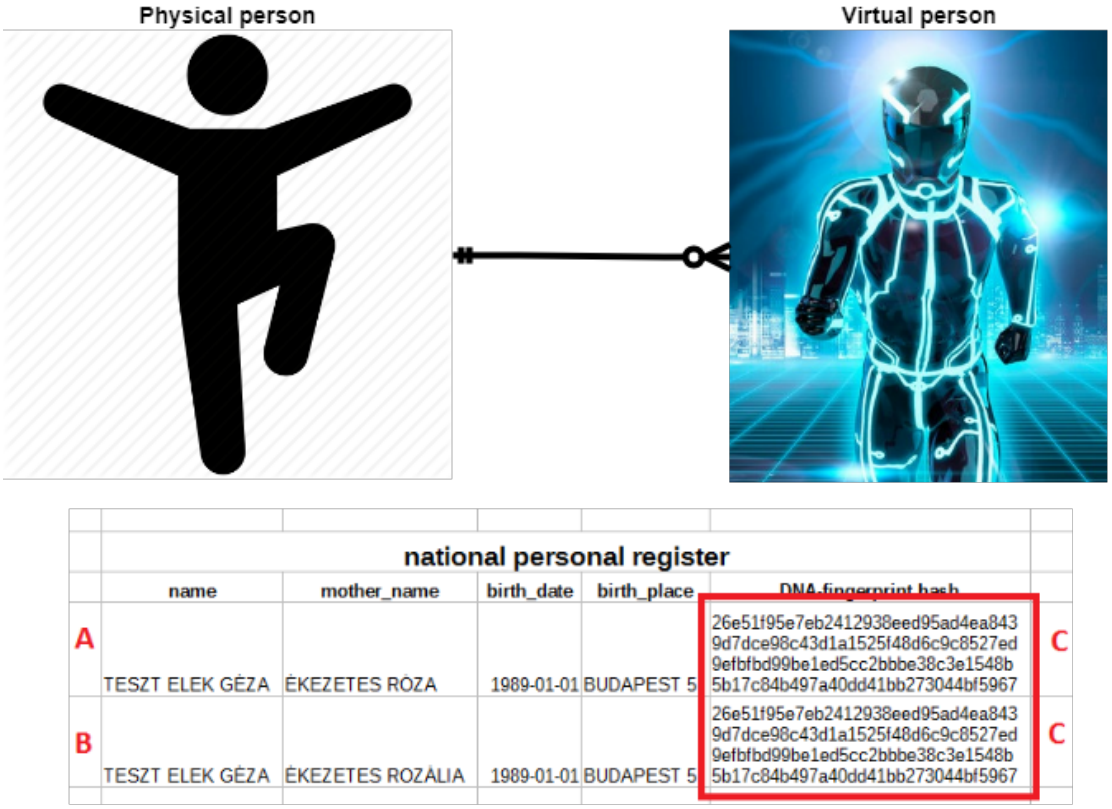

Figure 5 : A possible application of the national DNA-fingerprint databank (source: own)

In this example, we can use the DNA-fingerprint to prevent person duplicates. These duplicates will be created by various mistakes, for example missing accents, any attribute changed. Here we can see, one person of the national personal register will be in twice, different appearance in the databank because his mother's name was changed. In the real world, the person is the same but in the virtual world, the person is different the viewpoint of the system. This method has a limitation, the given DNA sample has to be clear from contaminations to give a real DNA-fingerprint which can be looked up in the national DNA-fingerprint databank. The other limitation is identical twins cannot be trusted identified based on DNA-fingerprint. The hash generation methods are the same as in the figure 4 model above.

6. Conclusion

Our goal in this research is developing a common national DNA-fingerprint system for identifying people and relating them to their digital data with a robust solution. We would not like to bring the DNA-fingerprinting technic as a new authentication method for users into public awareness. This is technically not feasible for fast authentication because the identification process is more than 90 minutes with the best devices too. Instead of the authentication method, it is a great choice to identify a person's identity in a trusted, widely available way. The other biometrics are really useful and easily usable but these are very damageable and not available for all. By these are not available for all, they divide people into different clusters based on the used biometric method, for example not all people can

give fingerprint sample, so the exceptions have to use another method such as the facial image. It would not be a problem in itself but with the very damageable property, a system based on these methods can be unused when a person lost its biometric which is related to itself.
The DNA-fingerprint is available for all and it is not damageable because if we see the extreme example when chopping a person, then its DNA almost always stays available. And if we move up one in the abstraction, we can realize that the DNA is available not only for the Human but for example for animals too, so we can use it not only for identifying Human entities with this method.
For the future, our plan is defining an exact method for identifying a person based on its DNA without any reason that causes people being clustered into more than one identification method, for example, identical twins' DNA-fingerprint equality or nowadays used fingerprint, facial image, voice.

Acknowledgment
We would like to thank our partners' helpful supports.
We would like to thank the Kocka Kör (English: Cube Circle) Talent Development Cultural Association and K + F Studio Ltd (R & D Studio Ltd., Hungary) for cooperation.

This work was supported by the construction EFOP-3.6.3-VEKOP-16-2017-00002. The project was supported by the European Union, co-financed by the European Social Fund.

References

[1] Jeff Desjardins. What Happens in an Internet Minute in 2018?. (URL: http://www.visualcapitalist.com/internet-minute-2018/)
[2] Nagy Ádám, Kölcsey Attila. Mit takar az alfa-generáció?. (URL: http://real.mtak.hu/62396/1/alfagen_metszetek_u.pdf)
[3] Erste Group. Erste Bank Introduces Video-Based Identification of New Customers. (URL: https://www.erstegroup.com/en/news-media/press-releases/2017/01/23/erste-bank-introduces-video-based-identification-of-new-customers-alias)
[4] 2015. évi CLXXXVIII. törvény. (URL: http://njt.hu/cgi_bin/njt_doc.cgi?docid=192456.348600)
[5] Molnár Csaba. 176 millió kamera kereszttüzében: a kínai Nagy Testvér mindenkire odafigyel. (URL: https://mno.hu/tudomany/176millio-kamera-kereszttuzeben-a-kinai-nagy-testver-mindenkire-odafigyel-2458366)
[6] Tóth Balázs. 60 000 fős koncerten is pontos a kínai arcfelismerés. (URL: https://index.hu/tech/2018/04/12/60_000_fos_koncerten_is_pontos_a_kinai_arcfelismeres/)
[7] Harrison Jacobs. China's 'Big Brother' surveillance technology isn't nearly as all-seeing as the government wants you to think. (URL: https://www.businessinsider.com/china-facial-recognition-limitations-2018-7)
[8] Biometrics definition. (URL: https://en.wikipedia.org/wiki/Biometrics)
[9] Alan Goode. Biometric Identification or Biometric Authentication?. (URL: https://www.veridiumid.com/blog/biometric-identification-and-biometric-authentication/)
[10] Danny Thakkar. Top Five Biometrics: Face, Fingerprint, Iris, Palm and Voice. (URL: https://www.bayometric.com/biometrics-face-finger-iris-palm-voice/)
[11] GDPR Regulation (EU) 2016/679. (URL: https://eur-lex.europa.eu/legal-content/HU/TXT/?uri=celex%3A32016R0679)
[12] ICAO 9303 Deployment of Biometric Identification and Electronic Storage of Data in MRTDs. (URL: https://www.icao.int/publications/Documents/9303_p9_cons_en.pdf)
[13] Fingerprint Technology Overview. (URL: http://www.identityone.net/BiometricTechnology.aspx)

[14] Jianjiang Feng. Fingerprint Recognition. (URL: http://www.comp.hkbu.edu.hk/wsb17/slides/Jianjiang_Feng.pdf)

[15] Fingerprint recognition. (URL: http://www.biometric-solutions.com/fingerprint-recognition.html)

[16] Yipeng Lu, Fari Assaderagh. 11.2 3D ultrasonic fingerprint sensor-on-a-chip. (URL: https://www.researchgate.net/publication/296872871_112_3D_ultrasonic_fingerprint_sensor-on-a-chip)

[17] Danny Thakkar. False Acceptance Rate (FAR) and False Recognition Rate (FRR) in Biometrics. (URL: https://www.bayometric.com/false-acceptance-rate-far-false-recognition-rate-frr/)

[18] Lutz Roewer. DNA fingerprinting in forensics: past, present, future. (URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3831584/)

[19] IntegenX RapidHIT System. (URL: https://integenx.com/rapidhit-system/)

[20] What is DNA?. (URL: https://ghr.nlm.nih.gov/primer/basics/dna)

[21] FBI Combined DNA Index System (CODIS). (URL: https://www.fbi.gov/services/laboratory/biometric-analysis/codis)

[22] Douglas R. Hares. Selection and implementation of expanded CODIS core loci in the United States. (URL: https://doi.org/10.1016/j.fsigen.2015.03.006)

[23] SHA3 documentation. (URL: https://keccak.team/specifications.html)

[24] Xiaoyun Wang, Yiqun Lisa Yin, Hongbo Yu. Finding Collisions in the Full SHA-1. (URL: https://link.springer.com/content/pdf/10.1007%2F11535218_2.pdf)

[25] Jacqueline Weber-Lehmann. Finding the needle in the haystack: Differentiating identical twins inpaternity testing and forensics by ultra-deep next generationsequencing. (URL: http://www.fsigenetics.com/pb/assets/raw/Health%20Advance/journals/fsigen/FSIGEN_monozygotic_twins.pdf)

[26] 1996. évi XX. törvény. (URL: http://njt.hu/cgi_bin/njt_doc.cgi?docid=26379.348026)

[27] ISO 3166-1 alpha-3 country code standard. (URL: https://www.iso.org/iso-3166-country-codes.html)

[28] ISO 8601 date standard. (URL: https://www.iso.org/iso-8601-date-and-time-format.html)

(All URLs were downloaded on 2018.11.24., last time.)