AARMS Vol. 14, No. 2 (2015) 149–159. **10.32565/aarms.2015.2.2**

Long-Term Storage of Digitally Signed Documents

GYURÁK Gábor[1]

Digital documents play an increasingly important role in our lives. Reliable digital storage of these documents is complicated and we have to deal with other problems if we would like to store these documents for a long time. Some documents, especially the most important documents are electronically signed. The long-term storage of electronically signed documents is more difficult, because we have to ensure the long-term validity as well. Electronic invoices (e-invoices) are also electronically signed documents and their role is becoming more important. The proposal of the European Committee on e-invoicing aims to facilitate the use of e-invoices. By 2020 e-invoicing will be general usage in the EU. This paper describes the problems in connection with long-term storage of digitally signed documents. Possible solutions are also presented. In connection to this, the regulation of preserving electronically signed documents is also examined from the point of view of Hungarian legislation. Finally it is shown how ETSI's (European Telecommunication Standards Institute) PAdES (PDF Advanced Electronic Signature) might support the long-term validity of e-documents, using the widely used portable document format (PDF).

Keywords: *long-term storage, digital signature, PAdES, pdf*

Introduction

Recording information has always been important in the history of humanity and also to save it for posterity. The oldest relics we found were cave paintings and were made about 13,000 BC, of which the most famous is located in Lascaux, France. In 3,000 BC a new age began with the formation of writing. After 5,000 years we are also able to recognize the ancient symbols of Uruk, and after 4,000 years we can easily read Hammurabi's laws and papyrus scrolls from the second century BC. We keep these several thousand year old relics in our libraries. [1]

The technological development in the last couple of decades has basically changed these thousand years old ancient traditions. Instead of clay tablets, papyrus, and paper we use magnetic tapes, optical disks, and other electronic devices.

New technologies have a great advantage over the old ones but we have to mention two non-typical properties. While the information was readable with human senses on conventional containers, like paper, now we need special devices to recognize the content of the new data storage (like a Blu-Ray Disc or a Pendrive). Another important difference is that while obtaining information from traditionally stored data does not require any special knowledge, interpreting and displaying data stored in binary format raises some difficulties. To better

1 University of Pécs, Faculty of Engineering and Information Technology, Pécs, Hungary, e-mail: gabor@gyurak.hu

understand these new challenges, let's think about a rightly famous Botticelli painting, made in 1486. Anyone who visits the Uffizi Gallery in Florence can enjoy this artwork. But to view a picture which is saved on a floppy disc and written in Dr. Halo's CUT[2] format may face some complications in opening it. The first problem is that, no one uses this old hardware nowadays, which reads the disc such as a floppy driver. Even if we can manage to read the disc, it will be hard to find software that can open such an obsolete file format and can display the picture.

Most information nowadays is published in digital formats. It makes creating, modifying and forwarding data much more convenient. This goes to the extent that even paper based documents are created electronically and then printed out. We can easily convert our older documents to digital format (e.g. by using a scanner) for easier access.

Digital formats are not only significant because of the more practical management of information, but for its preservation. Certain information such as pictures about unrepeatable events can be crucial for a person. In addition there is lots of information which we have to keep and protect for posterity. Examples range from scientific information to cultural heritage information, [2] but we can also classify the results of nuclear experiments in this group.

Regardless whether we speak about social or personal interests, there are data for which storage has to be guaranteed for decades or centuries. It is hard to explain what is meant by "long-term" data storage, due to the fact there is no clear-cut margin. Depending on the appliances, a few years can be classified as "long-term" but decade long storage can definitely be considered "long-term".

The aim of this paper is to give an overview about challenges with long-term digital signatures and also to describe one possible solution.

Long-Term Data Storage

Based on the previous discussion, the main problems of the long-term data storage of digital documents can be defined. The result of the advance of technology is that the hardware devices and software tools rapidly become out of date. We have to mention that nowadays data storage devices can only store data for a limited time. A commonly used optical data storage disk (CD, DVD, Blu-ray) can only store data for a few years in a trustworthy way. Sadly even the special coated, top of the line disks cannot be expected to work for more than 10–15 years. [3] By the way, more than 15 years storage with one disk is unnecessary because the technology becomes obsolete and there will not be any drives around to obtain information from the disks.[3]

There are only two ways to carry out the practical usage of the long-term data storage. The first one is the migration technique and the other is the emulation one. [4]

The main point of migration is that we transform our data to apply the new technology into a physical and logical frame. While using logical transformation, we convert from an old, obsolete format to a new, standardized one (e.g. converting a Word '97 document to a Word 2013 format). On a physical level while using migration, we have to switch over in certain periods to the new data storage technology (like when we are copying data from

2 Obsolete file format which was supported by the famous picture editor Paintshop in 1997.

3 There are newly developed disks, which according to the manufacturer's claims can store data for 1,000 years. One example for this is the M-DISC by Milleniata (www.mdisc.com)

floppy disc to CD), thus bypassing the technology becoming obsolete. Aside from switching to new technologies, we still have to consider the possibility of storing data in a traditional (non-digital) way (e.g. microfilm,[4] paper).

The other long-term storage solution is emulation, which saves our data in the original format. We eliminate these formats' obsolescence by eliminating the old system's hardware and software environment. There are several types of virtualization techniques to use for this purpose.

Authentic Documents in Electronic Format

In the previous section we talked about documents in general, however we have not mentioned their content. From the beginning of writing there have been "documents" containing vital information. Several techniques evolved in history that were supposed to protect these documents, mainly their confidentiality, integrity and authenticity.

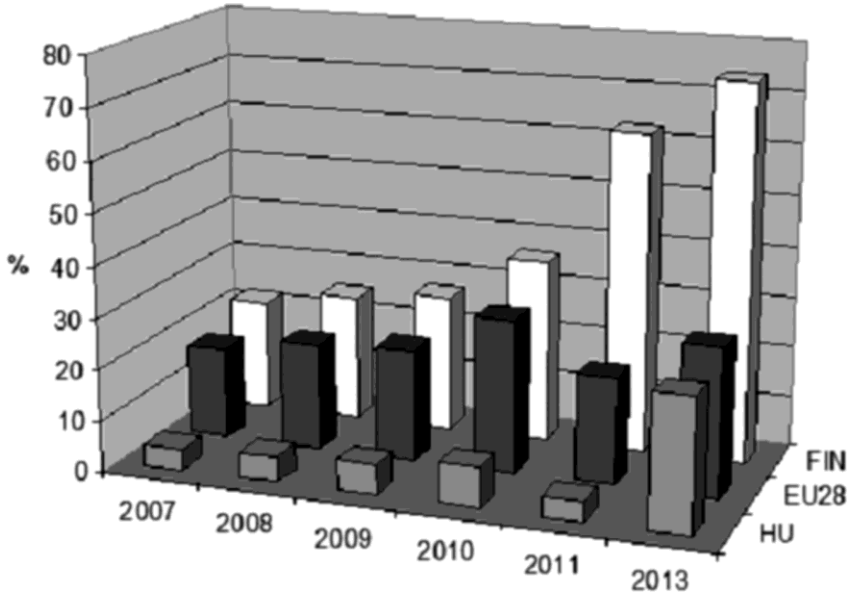
The purpose of confidentiality is to prevent non-authorized people accessing the content of the document. To achieve this, people have been using encryption ever since Sun-Ce's time.[5]

The authenticity of a document means that there is proof about who the creator of the document was. Traditional documents were marked by the handwritten signature of the creator. This is how a document became authentic. Most of the information nowadays is only available electronically, and it would be convenient to store authentic documents that way, as well. Nevertheless, in the field of authenticity, paper based documents are still more dominant, as the advantages of electronic documents (easy to create, copy and modify) in regards of authenticity quickly become their biggest flaws.

One of the important elements of the information society is the formation of a "paperless" government (e-government) system, and also the opportunity for electronic administration. Most of the procedures require the use of authentic documents, mainly electronic invoicing and electronic contracting. One of the core parts of "Europe 2020", the digital agenda classifies e-government as an essential part of a competitive union economy. The EU puts significant effort into spreading electronic administration, mainly towards making electronic invoicing a standard. In the European Commission's communication, titled "Reaping the benefits of e-invoicing for Europe" they called member states to make e-invoicing the standard way of invoicing by 2020. [5]

4 The oldest microfilm is more than 70 years old.

5 Sun-Ce was a military theoretician and mathematician in the 5th century BC.



Graph 1. The rate of eligible concerns of electronic billing. [6]

A statistic provided by Eurostat (Graph 1) shows how companies in different categories can provide the service of electronic invoicing. Hungary reached a huge breakthrough in the year of 2013 by catching up to the EU average. [6]

Electronic invoicing is showing a tendency of growth, and both parties are taking steps towards achieving the goal by 2020. Naturally, we would like to store authentic documents for a long time and sometimes our legal obligation is to store these documents long-term. Over the problems which we had met in the second paragraph, a lot of new challenges arise when we have to guarantee the long-term authenticity of documents. We are going to discuss this in later chapters.

Authentication in Electronic Documents

Technical Background [7] [8: 59–60] [9: 120–124]

The basic principle of authenticating electronic documents is essentially the same as with conventional documents: we sign the document and that signature identifies the creator of the document. The difference between authenticating electronic documents is that we use electronic signatures, generated with cryptographic algorithms.

Just as with handwritten signatures, digital signing should be done in a way that is verifiable and non-forgeable. That is, it must be possible to prove that a document signed by an individual was indeed signed by that individual and that only that individual could have signed the document. Let us consider, Alice and Bob[6] want to communicate via an electronic way. When Bob signs a message, Bob must put something on the message that is unique to

6 Alice and Bob are two commonly used placeholder names in cryptography.

him. Bob could consider attaching a MAC (Message Authentication Code) as the signature, where the MAC is created by appending his key (unique to him) to the message, and then taking the hash. But for Alice to verify the signature, she must also have a copy of the key, in which case the key would not be unique to Bob. Public-key cryptography is an excellent candidate for providing digital signatures.

The gist of the public-key cryptography system is that both parties get a pair of keys. One of the keys is a secret (private) key that the owner cannot give to anyone. The other one is a public key which can be accessed by anyone.

Suppose that Bob wants to digitally sign a document, *m*. We can think of the document as a file or a message that Bob is going to sign and send. To sign this document, Bob simply uses his private key, K_{Bpriv} to compute $E_{KBpriv}(m)$, where E is the encryption algorithm. This value is called the digital signature of the document. If Alice wants to verify the signature she has to take Bob's public key (K_{Bpub}) and she computes $D_{KBpub}[E_{KBpriv}(m)]$, where D is the decryption algorithm. It produces *m* which exactly matches the original document. Encryption and decryption are mathematical operations (exponentiation to the power of *e* or *d* in RSA). After this procedure Alice can be sure about the integrity and author of the message, because of the following reasons:

- Whoever signed the message must have used the private key, K_{Bpriv}, in computing the signature $E_{KBpriv}(m)$, such that $D_{KBpub}[E_{KBpriv}(m)] = m$.
- According to the main principle of the public key cryptography, the only person who could have known the private key, K_{Bpriv}, is Bob.

It is also important to note that if the original document, *m*, is ever modified to some alternate form, *m'*, the signature that Bob created for *m* will not be valid for *m'*, since $D_{KBpub}[E_{KBpriv}(m)]$ does not equal *m'*. Thus we can see that digital signatures also provide message integrity, allowing the receiver to verify that the message was unaltered as well as the source of the message.

One concern with signing data by encryption is that encryption and decryption are computationally expansive. Given the overheads of encryption and decryption, signing data via complete encryption/decryption can be overkill. A more efficient approach is to introduce hash functions into the digital signature. Hash algorithms take a message, m, of arbitrary length and compute a fixed-length fingerprint of the message, donated by H(m). Using a hash function, Bob signs the hash of the message rather than the message itself. Bob calculates $E_{KBpriv}[H(m)]$. Since H(m) is generally much smaller than the original message, the computational effort required to create the digital signature is substantially reduced.

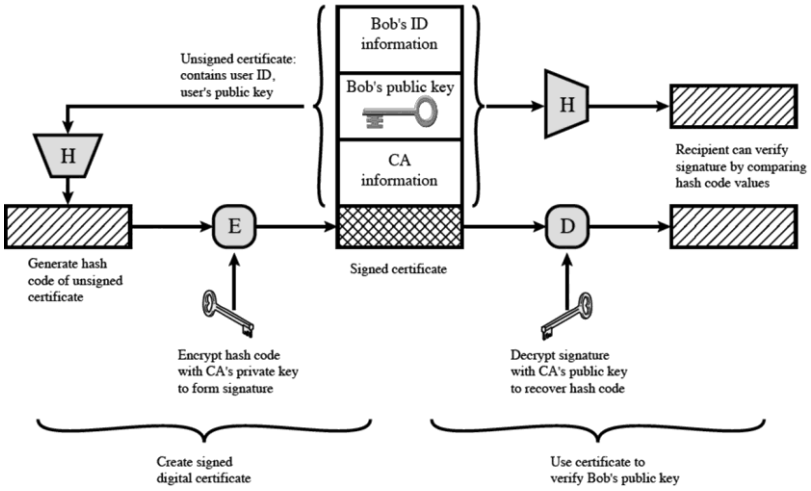


Figure 1. Public-key certificate use. [9: 60]

The digital signature system requires an underlying Public Key Infrastructure (PKI) with certification authorities. Public key certification is certifying that public key belongs to a specific entity. For example, when Alice wants to communicate with Bob using public key cryptography, she needs to verify that the public key that is supposed to be Bob's is indeed Bob's. Binding a public key to a particular entity is typically done by a Certification authority (CA), whose job is to validate identities and issue certificates. A CA has the following roles:

- A CA verifies that an entity (a person, a device, and so on) is who it says it is.
- Once the CA verifies the identity of the entity, the CA creates a certificate that binds the public key and globally unique identifying information about the owner of the public key. The certificate is digitally signed by the CA.

The user can then publish the certificate. Anyone needing this user's public key can obtain the certificate and verify that it is valid by means of the attached trusted signature. The certificate has an expiration date wherein the CA verifies that the particularly public key belongs to a user. The signature on the certificate is made with the CA's private key; therefore this can be verified with its public counterpart. The CA's public key is certified by another CA, thus creating a certificate-chain. The CA is responsible to publish Certificate Revocation Lists (CRL) containing information about invalid certifications and to make certifications verifiable online with Online Certificate Status Protocol (OCSP).

Legal Background

In the previous paragraph we introduced the technical side of the situation, which solves the document authentication in a technical way. However it cannot be used in practice, until it is acknowledged legally. The European Parliament realised the great potential of electronic signatures early, and in 1999 they provided member states with guidelines, with the directive 1999/93/EK. [10] Based on this directive, the law about electronic signatures (Esl – Electronic Signature Law) was passed in Hungary as well, in the form of the year 2001, XXXV. Law (hereinafter: Esl.); [11] it managed to provide sufficient legal background for electronic authentication

The law specifies four services:

- authentication service;
- timestamp service;
- device service;
- electronic archiving service.

The Esl. distinguishes qualified and non-qualified providers. Parallel to that we can talk about qualified electronic signature, increased security electronic signatures and other electronic signatures that do not fit either of those two categories.

Long-Term Certification Affected by Challenges

While signing an electronic document the signer takes responsibility for its content. When authenticating a document, we check whether its signature is valid or not. The Esl. only assigns legal consequences for documents with a valid signature. [11]

Steps of the authentication process:

- we create the hash print of the document H(M);
- we decrypt the signed hash print, using the signer's public key H(M)';
- if the two hashes match [H(M) = H(M)'], we can conclude that the signer of the document possessed the pair of the public key (private key).

With these steps, we can prove that the document has not been modified since it was signed and the signature was made by the private key that belongs to the public key. The next thing we have to check is who the set of keys belong to and whether or not the signer was the only one with access to the private key at the time of the signature. The focus is on the time of commitment, so it is very important that we inspect the circumstances at the time, as well, whether the validation happens right after the signing or decades later. The owner of the public key is verified by the certificate, the authenticity of which CAs are responsible for. The task is to verify if the signer's certificate was valid at the time of signing, as well as if there was a certificate-chain that could be traced back to a root CA's certificate and if all the elements in the chain were valid (the certificates were not suspended or revoked).

As we can see, validity checking is a very complex procedure, which makes inspecting a lot of data necessary. If all of this happens shortly after the signing, the validation process is relatively unproblematic. It is hard to actually say how long this period is exactly, but if we consider the standard expiration time of a certificate, then we talk about a 1–2 year long period (of course if the certificate is not revoked, in that case the time of the revocation is what matters). [12]

If the certificate expires or gets revoked, the signature still remains valid but validity verification becomes necessary. During the verification process, the following problems can arise.

The Signing Date

As far as we cannot prove the signing date, the signature will only be valid, if the certificate is valid too (this means, the validity time has not expired and has not been revoked either). We can increase the validation time, if we are able to prove the signing's date, i.e. putting a timestamp on it. From that point the validity of the timestamp will also be important for verification.

Revocation Information

A certificate validity can be suspended within the validity time, or also can be revoked, typically this happens when we suspect that the private key has been compromised. If there is a timestamp on the signature, in the case of the revocation of the certification, the validation can also been proved. If the signature was made before the revocation, it can be considered valid. The revocation information is published as a CRL by the CA, and it enables OCSP (Online Certificate Status Protocol) queries as well. [13] According to Esl., the service providers are liable to store data after the expiry of the certificate. They should store it for ten years. [14] This also means that if we want to ensure the validity for a longer term, then we have to collect the revocation information, and take care of their long-term storage.

CA Information

To establish the validation of the signature all the data in the certificate chain needs to be checked. The certification authority's certificates can expire. This question affects the validity of the timestamps, because time stamping is usually done by the same organization as the certification management. To achieve long-term validity, it is necessary to collect these data and store it.

Outdated Algorithms

In the background of the electronic signature there are certain cryptographic procedures, encryption algorithms (e.g. RSA [Riverst Sharim Adleman] algorithm), hash algorithms (e.g. MD5, SHA–512) to operate. These have properties that allow the system to work safely, meaning that with the current level of technology, there is no efficient way or sufficient computational power to compromise the system. According to our current knowledge, there is no appropriate way for integer factorization. [7] This is what the RSA encryption is based on, and this is why RSA based electronic signatures are considered safe.

We have arrived at yet another point, where we have to pay attention to the time factor. Those algorithms that we use today might become obsolete in a few years, but decades later will be surely outdated.

The MD5 hash algorithm could be a great example how cryptographical building blocks become obsolete. [15] It was widely used before the millennium. MD5 hashes of documents were provided with an electronic signature. One of the criteria of a hash algorithm's usability is that it has to be collision resistant. It means that it is hard to find two messages that have the same hash print. As it turned out, the MD5 does not meet these criteria so it cannot be used for cryptographic applications.

Another widely used algorithm is the SHA–1, [16] which is not allowed to be used for cryptographic purposes for CA organizations in Hungary since 31 December, 2011.

With the developments of technology a lot more computational power is possible, which makes a brute force attack on one of these solutions really easy. Fortunately, the regulatory side recognized this vulnerability and now there is legislation in place to make companies use safe algorithms and appropriate long keys.

Solutions

We have to find a solution to the problems presented in the previous chapter. A solution that can guarantee the long-term validity of electronically signed documents in technological and legislative aspects.

The 4th paragraph of PadES developed by ETSI, the PadES-LTV (Long-Term Validity) is a development that extends the PDF format with capabilities that allow the long-term validity of an electronically signed document. [17]

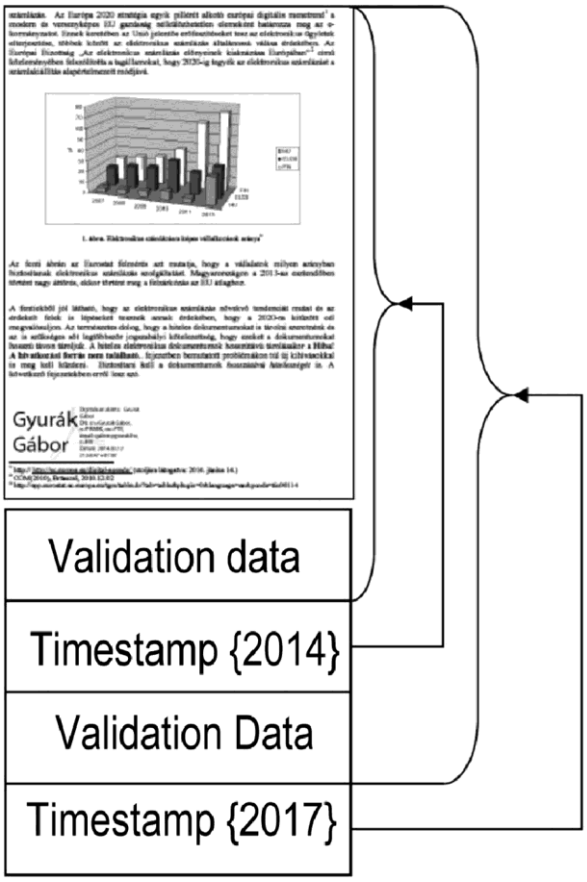


Figure 2. PaDES document. (Created by the author)

The base is a PDF document, provided with an electronic signature and timestamp. Then an extension is added to this, containing additional data that we use for validating the signature (validation data):

- certificates of CAs in the certification chain;
- revocation information (in the form of CRLs and/or OCSP answers);
- certificate of timestamp provider.

According to Figure 2, a timestamp is added to this which provides the document's validity within its expiration date even if the signer's certificate had already expired. The signature can be verified even if the CA's information and revocation information are not available, as they were attached to the document.

If we want to store the document long-term, we can ensure validity through repeating the previously mentioned steps (attaching validation data and timestamp). All we need to pay attention to is that the "update" has to happen before the timestamp's certificate expires and the new algorithm used at the time stamping must be up to the current standards (secure algorithm and long keys). [18] [19: 7–8]

In addition to the technical solution, the system only works if it is supported with the appropriate legislative background. In Hungary, the 114/2007. (XII. 29.) MET (Ministry of Economy and Transport) ministry decree [20] about digital archiving regulates the long-term storage of digitally signed documents.

The decree's para 4 (4) distinguishes, the obligatory period of time to preserve the document, long-term storage, that according to the law in place, means more than 11 years. In this case, it is the job of the one in charge of preservation, to:

- take care of the acquisition and preservation of the information necessary for electronic signatures long-term validation;
- place a timestamp on the certificate-chain, provided by a qualified provider;
- repeat the previous step if the cryptographic algorithms used become obsolete.

We can meet these regulations ourselves, or we can hire an archiving provider to do the job for us. In the latter case, we have to assume the provider does his job well, so in case of a dispute, the conflicting party has to prove the problems with the document's authenticity.

According to the law about accounting, the obligatory time period to preserve electronic invoices is eight years. Based on the rules of archiving this does not classify as long-term so the strict regulations do not apply here. Although, from a technical standpoint, it would still be justified to use the archiving methods mentioned above, even in this "short" period, as the problems outlined in the earlier paragraph can affect our documents in this period, as well.

Summary

From the previous chapters we can clearly see that long-term data storage, especially of authentic documents is a difficult and expensive job. With the above mentioned technologies, we can guarantee long-term authenticity but we should not forget about the usual flaws of long-term storage. Aside from guaranteeing authenticity, we have to still make the interpretation of the original document possible thus reaching a point where the questions discussed in the second chapter come up.

According to the decree of 114/2007. (XII. 29.) para 2 (2):

"The one bound for preservation has to guarantee that the documents stored remain readable – through supplying the appropriate software and hardware environment to open the document – for the time period of said preservation." [20]

According to what we have mentioned above, the migration technique can not be used on a logical level, as the document has to be kept in its original format. The solution to this problem is the emulation technique, or preserving the original hardware and software environment.

References

[1] VÁRKONYI N.: *Az írás és a könyv története.* Budapest: Széphalom Könyvműhely, 2001. [2] ZHOU, M., GENG, G., WU, Z.: *Digital Preservation Technology for Cultural Heritage.* Berlin: Springer, 2012.

[3] PEEK, H., BERGMANS, J. HAAREN, J. van, TOOLENAAR, F., STAN, S.: *Origins and Successors of the Compact Disc*. Berlin: Springer, 2009.

[4] BORGHOFF, U. M., RÖDIG, P., SCHEFFCZKY, J., SCHMITZ, L.: *Long-Term Preservation of Digital Documents*. Berlin: Springer, 2006.

[5] EUROPEAN COMMISSION: *EUROPE 2020.* http://ec.europa.eu/europe2020/ (downloaded: 17 12 2014)

[6] EUROSTAT: *Eurostat Statistics Graph: Enterprises Sending and/or Receiving e-Invoices.* http://epp.eurostat.ec.europa.eu/tgm/table.o?tab=table&plugin=0&language=en&pcode=tin00114 (downloaded: 17 12 2014)

[7] BUTTYÁN L., VAJDA I.: *Kriptográfia és alkalmazásai.* Budapest: Typtotex, 2012.

[8] STALLINGS, W., BROWN, L.: *Computer Security Principles and Practices.* London: Pearson, 2012.

[9] SOLOMON, G., CHAPPLE, M.*: Information Security Illuminated.* Burlington: Jones and Bartlett Learning, 2005.

[10] *The European Parliament and the Commission's directive of 1999/93/EK about electronic signatures social program, 13 December 1999.*

[11] *2001.* évi *XXXV. törvény az elektronikus aláírásról.* http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A0100035.TV (downloaded: 07 01 2015)

[12] SONG, S., JAJA, J.: *Techniques to Audit and Certify the Long-Term Integrity of Digital Archives.* Berlin: Springer, 2009.

[13] BERTA I.: *Nagy e-szignó könyv.* Budapest: Microsec Kft., 2011. [14] *2001. évi XXXV. törvény az elektronikus aláírásról*.

[15] *RFC 1321: The MD5 Message-Digest Algorithm.* www.ietf.org/rfc/rfc1321.txt (downloaded: 05 01 2015)

[16] *RFC 3174: US Secure Hash Algorithm 1 (SHA1).* https://tools.ietf.org/html/rfc3174 (downloaded: 05 01 2015)

[17] *ETSI TS 102 778 (2009-07) Electronic Signatures and Infrastructures.* www.etsi.org/deliver/etsi_ts/102700_102799/10277801/01.01.01_60/ts_10277801v010101p.pdf (downloaded: 12 01 2015)

[18] POPE, N.: Protecting Long-Term Validity of PDF documents with PAdES-LTV. In. POHLMANN, N., REIMER, H., SCHNEIDER, W. (Eds.), *ISSE 2009 Securing Electronic Business Processes,* Berlin: Springer, 2009. 320–327.

[19] BLANCHETTE, J.: The Digital Signature Dilemma. *Annales des Télécommunications,* Mai/Juin 2006, 1–18.

[20] *114/2007. (XII. 29.) GKM rendelet a digitális archiválás szabályairól.* http://net.jogtar.hu/jr/gen/hjegy_doc.cgi?docid=A0700114.GKM (downloaded: 07 01 2015)