Information Society, E-Records and the New Archival Science

Lajos **KÖRMENDY, *Ph.D.***
former National Archives of Hungary
e-mail: kormendy.lajos@gmail.com

Information Society, E-Records and the New Archival Science

ABSTRACT

The study shows how information society is reflected in e-records (volatility, dynamism, mixed private and public spheres), how it transformed the records themselves (no physical aspect, complicated structure, separated elements), and how it conquered the human memory. The basic dilemma of long term preservation is that e-records are dependent upon hardware and software which change rapidly, therefore we are unable to preserve them unchanged. The study talks shortly about integrity and authenticity of e-records which are basic requirements, and then outlines the major preservation strategies (migration, emulation, technology preservation, post-custodial archives, cloud computing) used by archives. In the digital world archival science adopted a lot from information science which caused that standards (ISADg, EAD, OAIS) are the new milestones of archival science. The greatest „yield" of this cooperation is that the data management and procedures are automatable. However, archivists must keep their classical synthetic work in digital world, too: to know and describe the record creator agency, its economic, political and societal environment, to build the system of fonds, etc.

Key words: e-records, long term preservation, archival science

La società dell'informazione, i documenti elettronici e la nuova archivistica

SINTESI

Lo studio mostra come la società dell'informazione si rifletta nei documenti elettronici (volatilità, dinamismo, sfere miste pubbliche e private), come ha trasformato il documento stesso (nessun aspetto fisico, struttura complessa, elementi separati) e come abbia conquistato la memoria umana. Il dilemma fondamentale della conservazione a lungo termine è che i documenti elettronici dipendono dall'hardware e dal software che cambiano rapidamente, e pertanto non siamo in grado di preservarli invariati. Lo studio parla brevemente dell'integrità e l'autenticità dei documenti elettronici, che sono requisiti fondamentali, e quindi delinea le principali strategie di conservazione (migrazione, emulazione, conservazione di tecnologia, il cloud computing) utilizzato dagli archivi. Nel mondo digitale la scienza archivistica ha preso molto dalla scienza dell'informazione, il che ha prodotto standard (ISADg, EAD, OAIS) che sono le nuove pietre miliari della scienza archivistica. Il più grande risultato di questa cooperazione è che la gestione dei dati e le procedure sono automatizzabili. Tuttavia, gli archivisti devono anche conservare il loro classico lavoro sintetico nel mondo digitale: per conoscere e descrivere il soggetto produttore, il suo ambiente economico, politico e sociale, per costruire il sistema dei fondi, ecc.

Parole chiave: documenti elettronici, conservazione a lingo termine, archivistica

Informacijska družba, e-gradivo in nova arhivska znanost

IZVLEČEK

Študija prikazuje, kako se informacijska družba odraža v e-gradivu (nestanovitnost, dinamičnost, mešana zasebna in javna sfera), kako se je preoblikovala gradivo (nobenega fizičnega vidika, zapletena struktura, ločeni elementi), in kako je osvojila človeški spomin. Osnovna dilema dolgoročne hrambe je, da je e-gradivo odvisno od strojne in programske opreme, ki se hitro spreminja, zato ga ne moremo ohraniti v nespremenjeni obliki. Študija govori tudi o celovitosti in avtentičnosti e-gradiva, ki predstavljata osnovne zahteve, nato pa opisuje glavne strategije hrambe (migracija, emulacija, ohranjanje tehnologije, hramba v oblaku), ki jih uporabljajo arhivi. V digitalnem svetu je arhivistike veliko prevzela od informatike, kar je povzročila, da so postali standardi (ISADg, EAD, OAIS) novi mejniki arhivistike. Največja pridobitev tega sodelovanja pa je, da so postopki upravljanja s podatki in špostopki avtomatizirani. Vendar pa morajo arhivisti tudi v digitalnem svetu obdržati svojo

klasično sintetično delo: poznati in opisati ustvarjalce arhivskega gradiva, njihovo gospodarsko, politično in družbeno okolje, da bi lahko izgradili sistem fondov itd.

Ključne besede: e-gradivo, dolgoročna hramba, arhivska znanost

1 Society and archives, challenges and responses

At the end of the 19th century, the development of the industrial and democratic societies required that public and private records be kept and retrievable. The response of the archivist's community to this challenge was the formulating of the concept of the principle of provenance which enabled archives to collect and process the records of agencies and prominent people in a comprehensive way, and the users to find relevant information in a mass of documents, as well as the authenticity of the records could also be ensured.

In the mid-20th century, due to the more and more expanding state involvement in the everyday life of society, the quantity of the created records increased rapidly which made the situation no more manageable with the old methods. Then archivists developed new approaches of the records' evaluation and disposal (see Schellenberg, sampling), they started microfilming and cooperating with records managers in order to get records of better quality from the agencies.

Nowadays the e-records make an even greater challenge to archival science: everything is questioned, new approaches are needed from the concept of the record to the preservation issues.

The e-records are the products of a society, which has dramatically changed in the last few decades, and what we call nowadays information society. IT has penetrated into all aspects of human life and transformed it, especially the human communication. The quantity of information increases faster and faster and more and more people appear in the web and leave traces (millions of private web pages, comments, blogs, etc.) which change rapidly. The function and the style of e-mails, blogs, instant messages and comments resemble much more a (phone) conversation than a traditional written information whose formula were developing in centuries. The use is very casual, see for example the "new orthography", or the great variety of attached documents (effects, images, sound files), or the disappearance of elements like form of address or subject. Earlier we resigned ourselves to that spoken information were unarchivable, nowadays we endeavour to keep digitized verbal records.

In the world of the web, the majority of the e-records is transitional or dynamic; there are more and more daily changing (increasing) databases. It is hard to separate civil and official activities, the private and official mails are often mixed in the mailboxes of the staff, which causes headache for records managers and archivists who want to archive e-mails. At the very start the web pages were mostly the communication tools of companies and organizations, nowadays private sector also uses it extensively. Community sites progressed in an opposite way: at the beginning first of all individuals or civil groups used them, later public institutions and public figures like politicians also discovered them.

IT has gradually conquered the human memory. At the beginning big registers and catalogues had been "mechanized", then came textual documents, tables and drawings. The next stage was the digitization of (both still and moving) images as well as sound records, and owing to the 3D printers, which appeared only a couple of years ago, now we are able to digitize physical objects, too. Nowadays we can say that IT enables us to create or reproduce the majority of the objects of the human memory. The outcome is the explosion of the quantity of electronic data, which is a self-generating process.

The shift from analogue to digital literacy as well as information fixing, which started in the second half of the 20th century, had two driving forces: to compress and to manage better information (by arrangement and retrieval). However, to read digitized (coded) information technical tools (hardware, software and IT environment) are needed, not only for magnifying but also for understanding; otherwise, information would be incomprehensible for human being. The trustworthy long term keeping of records that we know in analogue world is impossible in the digital world, and the technical reasons are well known: 1) the technical tools which are necessary to create, manage (read) and keep e-records determine the e-records; 2) since the tools change rapidly over time it is impossible to keep

the e-records in their original form. Owing to hardware and software, obsolescence e-records will be at risk after 5-10 years.

However, the trustworthy long term keeping of records has archival conditions, too, these are less known, and if we do not respect them, they can cause serious problems. Every alteration originates from four facts: 1) the e-record has no physical aspect; 2) its structure differs from the analogue one and strongly depends on the technology; 3) it is easy to create, alter, multiply, combine and disseminate the e-records; 4) because of the above-mentioned reasons the quantity of e-records increases rapidly. The consequence is that IT has deeply transformed the terms, the techniques and the rules of archival science. Let us see first how IT has transformed the records themselves.

2 The constituent elements of the archival records

The archival record, in general, is composed of six elements: support, form, content, identifier, structure and context[1].

Support

In the analogue world, we consider support that physical object (parchment, paper and ink, microfilm roll, etc.) which "carries" information and makes possible its appearance. In that world, the "carried" information can be read with naked eye. In the electronic era, the concept of support has radically changed first of all because the e-record has no physical aspect. This recognition has become evident owing to the spread of internet: an e-record can be kept anywhere in the world, on any kind of support (hard disk, CD, pendrive, memory of a computer, etc.), in any allocation (e.g. dispersed in several files). Furthermore, the existence of the bits representing the digital information is not enough in itself to visualize it, i.e. "to carry" the information; we also need technical tools (hardware, software and IT environment) for this. This means that, in broad sense, the support of e-records includes the technical tools, too, which are partly physical and partly virtual. Contrary to the analogue world, the sameness of the support is not a condition of the edition and the preservation of the e-records, for this reason the role of the support has considerably decreased in the digital world.

Form

Properly speaking the form means how the record looks like: in case of written document what colours, size, characters, figures, letterhead, etc. it has. In the analogue world, the form is inseparable from the support, and we automatically notice it. In the digital one it is determined by the given hardware and software (e.g. word processor or database management system) as well as by our decision on typeface, letterhead, colours, etc. Hardware and software display the document we are reading according to exact specifications (metadata), and if full compatibility is not ensured, e.g. the word processor has not an appropriate character set, the form changes.

The form can be integrated with the content or can be in a separate (stylesheet) file, so we can have several forms belonging to the same content, this can occur in XML files. As form separates from the content and strongly depends on hardware and software, that change by user and over time, its importance has decreased in the e-world; to be more precise, archivists must often reach a compromise when they are unable to preserve the exact form of the records. For instance, it happens sometimes that when converting the files we have to give up some form elements. By the way, the importance of the form varies by document type, e.g. it can be important in case of a graph but the character set of the hit list of a database is generally not considered as essential.

Content

Content is the immanent information relating to the subject of the record. The function and the importance of the content did not change in the digital world, because it is immaterial, and this coincides perfectly with the substance of the digital world. We must remark that both form and context (see below) may have content connotation, but since these are not necessarily immanent or relate to the subject, we do not treat them here.

1. There are other divisions, too, e.g. InterPARES project distinguishes, on the basis of diplomatics, four basic elements: documentary form, annotations, medium and context. See (InterPARES 1)

Identifier

The identifier has a twofold role: 1) it identifies uniquely the record, preferably in a way that one will not change it in the future; 2) this unique identification makes possible to assign the record's exact position in the system and to show the aggregate of records i.e. how the records are connected with each other. The unique and permanent identifier is an important requirement to keep record's identity and authenticity.

The importance of the identifiers has increased in the digital world. On the one hand, to connect databases, for example to integrate digitized records in an archival portal is a daily practice, and in such cases, the lack of appropriate identifiers would cause serous troubles. This is why that there are international projects to determine the rules of standard and unique identifiers[2]. On the other hand, as we will see below, the structure of the e-records is much more complicated than that of the analogue ones, for this reason they may contain such identifiers, which are invisible for the users, and which were unknown in the analogue world. The records of a big database can be distributed in a hundred of data tables, and the database software can "keep order" if all data are provided with appropriate identifiers. To have these invisible identifiers is indispensable requirement for converting the records. Moreover, there are also technical identifiers, which describe the IT environment of the records (files), e.g. its format. (Such identifiers are contexts in another approach.)

We mostly treat identifiers as metadata, and they can be integrated with the content and the form into a file, or sometimes into the file name, but they can also be collected into a separate table.

Structure

As the records are complex from several (content, formal, logical and IT) aspects, they have structures. Diplomatics, archival and information sciences distinguish different record structures, now we are speaking about the archival one, which is the "carrier" of the coherence and the integrity of the record, and what is indispensable for the long-term preservation. A record has inner and outer structures, and the aggregates of records are assembled owing to the latter one.

- Inner structure

  Information and data structure into document/record: it is enough to look at a traditional letter, and we can see and understand the addressee, the subject, the data and the signatory, i.e. the inner structure. In case of an e-mail when we fill the field of addressee or subject, the software generates automatically this structure by metadata with the help of which the reader's computer will be able to display the same layout. In the paper world, we can visualize the structure of a document or an aggregate of records by highlighting the titles and subtitles; in case of a web page, we must indicate this to the computer by tags.
  In case of e-records, technical data may complete the above-mentioned structural data, which did not exist in the analogue world or they were not considered as significant ones, for instance information concerning the transmission path and the exact reception time of an e-mail.
  The inner structure of the e-records are in danger especially when converting: the loss of significant metadata may cause strong degradation of the record's value.

- Outer structure

  The archival records are connected with each other based on content, procedure, function or form, so the documents belonging to the same business form a file; the aggregate of files of same type is called sub-series; the sub-series create a series, etc. According to this, we can distinguish level of records, files, sub-series, series, sub-fonds and fonds (ISADg). The relations between the units (aggregates) usually correspond to the filing plan of the record creator, which more or less reflects its functioning.
  The outer structure does not stop at the fonds level. An archival institution gets records from

2. ISADg, the first international descriptive standard (1994), also stressed the importance of the unique identifier. In the digital world there are a number of standards of this kind in order to assure that a digital object can be found even if its location has changed. See e.g. Universal Resource Names (URN) or Digital Object Identifiers (DOI) or Persistent Uniform Resource Locators (PURL).

many agencies, which means that it must integrate many fonds into a big structure otherwise it would be unable to build and manage a coherent registry and retrieval system. This system over the fond level can be called macrostructure. However, individual records may also be connected to records belonging to different aggregates. The best example for this is a web page, which can be connected with many other web pages by links.

In the outer structure, the direction of the connections can be vertical or horizontal. The adjective vertical refers to hierarchical connections between aggregates (e.g. sub-fonds, fonds, group of fonds). The adjective horizontal refers to "equal" connections, e.g. which series form a sub-fonds, or an e-mail is a response to an earlier message or it is in connection with a couple of paper records

To keep the information value of the records we need to preserve the outer structure. For instance if we disperse the records of a series, and we do not indicate that they are connected, their information value decreases significantly, despite their physical intactness, because we cannot attach them to the reason of creation and the business they belong to as well as the record creator. This is why it is important to reveal, document and keep the structure, in the analogue world the catalogues, finding aids and other descriptions serve this purpose.

The digital world is more complicated because e-records have typically much more external connections: the web pages are full of links, the e-mails have many attachments. On the other hand, the simple fact that the analogue, three-dimensional world is reduced to two dimensions is a considerable handicap, which must be compensated by structural metadata. When we work with traditional paper records we automatically recognize the boxes and the files in folder - parts of the outer structure - because they are separated from other boxes and folders, and we can turn and arrange the pages for example in logical order. The displayed e-records are aligned in sequential order, and it is much more difficult to handle them. In order to compensate a disadvantage a database (i.e. an aggregate of metadata) is often assigned to the e-records which shows the connections and the separations, and help handling, browsing and searching the records.

Context

The professional literature calls context the data and information concerning the record creator (history, activity, functions, organisational structure, relations with other agencies or individuals, legal and political societal information) as well as relating to the record itself[3]. These information and data are often implicitly in the records, for this reason, archivists must expose them; sometimes they have to be completed with data from other sources. The archival context is typically reported as metadata when describing the records.

However, e-record is a digital object, too, thus it has significant technical attributes, such as data relating to file format, which are vital for future use and preservation. The technical context, which describes the IT environment, must be accurate because every error can cause a failure. Due to the technical requirements, the importance of the context has increased in the digital world.

In summary, we can say that in the analogue world there was a consensus that the support, the form and the content represent together the record, and they are inseparable. This approach weakened a little bit by the spread of analogue copies, namely electrostatic (direct) copies and microfilms, when new supports emerged. In the digital world, everything has radically changed: the records have been dematerialized, the consisting elements have transformed, they have been separated from each other and completed by technical attributes, their role has increased or decreased. All the elements which are to make order in this complicated system such as identifier, structure and context, have become more important, on the other hand, we will see it later, they are indispensable for the record's identity, integrity and authenticity. These changes have made records more complicated and vulnerable - let us think over what happens if one of the elements is lost because of data corruption -, and this is why that their preservation is so difficult.

3. You can read a good definition of archival context in *ISAD(G).* InterPARES distinguishes five sorts of contexts: juridical-administrative, provenancial, procedural, documentary and technological ones. See (Duranti &Thibodeau, p. 18)

3 Metadata

I would like to say just a few words about metadata, which are absolutely essential in the digital world. The common definition of metadata is "Data about data" (in our case: data about records). Metadata were known in the analogue world too, we called them for example library catalogue or archival finding aid. However, their role has considerably increased in the digital era because most of the record's elements appear as metadata, so we can refer to the support by (technical) metadata, the identifier is also metadata, and we describe the structure and the context by metadata. They both belong to the digital born records and the scanned images. In case of images and sound records metadata are particularly important because contrary to the textual documents we cannot search for words, thus the only tools of searching and registering are metadata.

A basic characteristic of metadata is that, although they can be made human readable, their most important function is to be processable by machines (Day, 2005, p.12).

4 The integrity and the essential characteristics of the e-records

Integrity means completeness and coherence. It is easy to check the completeness and the coherence of a traditional record, what is not the case with e-records because they are much more complex, and their elements can be separated: it is much more difficult to assure and check their integrity although it is vital for their authenticity, trustworthiness and usability. What is more this integrity must be assured not only in single records level but also in case of aggregates of records.

A substantial condition of integrity is that we have all constituent elements of the record. However, as we saw above the support may be different - no matter that we download the e-records from hard disk or CD -, sometimes the form can alter, but the content, the identifier, the context and the structure must be identical with the original. In the digital world, the structure "carries" first of all the coherence of the records, and both the inner and the outer structure are in danger particularly when converting: metadata (e.g. addressees of e-mails) may unnoticeably disappear without which the information value and authenticity of the records decreases significantly. The more the inner and the outer structure of a record or a record's aggregate is complex the more difficult is to keep their integrity. Databases, as records, have the most complicated structure, although the user cannot see it, for instance the computer can gather the data of a list displayed on the screen from many data tables.

The international literature deals a lot with the integrity of single records, i.e. the inner structure, but barely deals with the outer one, i.e. the integrity of aggregates of records. However, it would be an important issue, because as we saw above single records do not stand in themselves, their value and purport prevail only together with the connected records, thus the outer integrity must be kept, too.

The outer integrity must be ensured in a different way as the inner one. The majority of the records created - both the analogue and the electronic ones - must be destroyed (disposal) before transferring them to the archive, and this is against the outer integrity. However, if the selection is done in an appropriate way, in accordance with the records schedule, cutting the least connections, and if it is documented properly, then we can ensure not only authenticity but outer integrity of the records, too. However, we must acknowledge that we are unable to ensure the outer integrity of certain type of records, such as web pages, which are full of links connecting different websites. On the one hand because the web pages change frequently (they are dynamic), on the other hand because the pages connected contain also links pointing to other websites. It is impossible to archive all the sites connected because they number increases in geometric progression.

The National Archives of Australia recommends that we preserve first of all the record's essence, such as the content, the font type and size and the layout (form) in case of word-processed documents, but not the toolbars or the button functionalities. The Australian archivists recognize that "Determining the essence of records is not a science and is open to subjectivities and archival interpretation..." For this reason they determine the essence of all types of record they have (word-processed document, email, etc.), then they can apply optimal preservation treatment with any record (Heslop&Davis&Wilson 2002, pp.13-15).

5 The authenticity of the e-records

The literature written about the authenticity of the e-records may fill libraries, here I can only give a short summary. According to a common definition, a record is authentic if it is what it purports to be. The authenticity of the e-records is mostly doubtful because they can be easily manipulated without leaving a trace, for example, we can easily alter a sentence in a text file, discolour an image in an image file or erase the background noise in a sound file. And what is more, as we said earlier, the separation and the vulnerability of the record's constituent elements, the changes of support and the alterations of the form are all against the authenticity.

The issue of authenticity has been studied in detail by the InterPARES project. According to this authenticity must be ensured on the level of

1. the records by their identity and integrity
2. the record creator and preserver by

- controlling record transfer, maintenance and reproduction
- access, security and preventive procedures against loss and corruption of records as well as technological support
- appropriate procedures for disposal
- documentary forms of records associated with the procedures (InterPARES 2, pp. 4-10).

6 The major strategies of e-records' long term preservation

Nowadays we can state with certainty that we are unable to preserve the e-records unchanged for decades or centuries, or rather, if we do so, they will be unreadable after some time. The most we can do is to keep the records in a condition that we will able to reproduce them in the near future[4]. Every institution responsible for long-term preservation of records must elaborate its preservation strategy, i.e. those basic methods by which it can keep this capability of reproducing.

In the last few decades archives tried several strategies of preservation, some of them have been successful, i.e. proved themselves to be durable solution, we will talk about them below, others have been unsuccessful. Although the latters can be considered as failures, it is worth mentioning two of them shortly. The first is called technology preservation, which means that the archive keeps not only the original records but also hardware and software, which manage them. Nowadays this kind of strategy has almost disappeared because maintaining and managing an ever-growing hardware and software collection would cause an unsolvable problem after a few decades. On the other hand manufacturers stop producing obsolete parts after a decade, thus the machines (and the records managed by them) live as far as the parts are operable. Understanding and handling several hundreds of software would also make an insoluble problem for the archives' staff. Nevertheless, in some cases, there is a raison d'etre to apply technology preservation: it is recommended to keep some obsolete hardware unit for a couple of decades that we may need in the future. (For example, a floppy disk driver, which can read floppy disks, transferred to the archive, and so it will be possible to copy the records to a standard support).

The so-called post-custodial archives was a popular issue in archival conferences in the nineties, at the beginning of the Internet revolution. According to this theory, the future archival institutions will not preserve e-records, the producers (agencies) will keep and maintain them (and convert them if necessary), because in the era of Internet it does not matter where (in which computer) the records are. According to this theory, the archives' task will be only to keep intellectual control (description, access, etc.) over the records[5]. The post-custodial strategy has not spread because the record creators have not been interested in investing money and taking pains to toil for the archives for centuries. However the cloud computing, what is now quite new matter, may remind us of, in some respects, the post-custodial strategy.

Cloud computing means that the records are kept in a remote server of an IT company which

4. Kenneth Thibodeau, quoted by (Eastwood, 2002, p. 77)

5. A summary of this theory see (Cook, 1996, pp. 205-207). This strategy was applied by the National Archives of Australia in the nineties. See (National Archives of Australia, p. 6)

provide the archive with not only storage but also hardware and software operation. The access of the records is done through Internet. The archives must do all professional work (elaborating strategies, disposal, describing the records, establishing a retrieval system, etc.) The cloud can be advantageous from various aspects - for instance, it can be cost effective and the archive does not have to worry about infrastructure and IT experts - but it has a considerable risk: the records are out of the control of the archive. Although in the developed countries the security level of the cloud is high, every archival institution which considers applying cloud computing must carefully calculate the pros and cons. If the answer is positive, the archive must build up several security elements such as appropriate export functions (in order to retrieve data from the system safely and easily) or making multiple copies.

Migration

Most of the archival institutions archiving e-records apply the strategy of migration. Migration means that the archive converts the files from obsolete formats to new ones. When selecting the new format several criteria must be taken into consideration such as how widespread it is, whether its specifications are published or not, how it depends on technology, etc. Since formats, also change over time migration must be done repeatedly (but as rarely as possible).

During migration data, loss is inevitable, but it does matter the quality and the quantity of the data lost. We must not lose data representing content, metadata and important functions, or anything that belongs to the record's essence mentioned above. However, as regards data relating to editing on the screen we can be more permissive. Consecutive migration (when we migrate the records already migrated earlier) may cause cumulative data loss therefore it is to avoid (for example we can always make conversion from the original files) or to do as rarely as possible (for example we make conversion from long lasting master files). In order to control data loss it is not enough to apply a random control procedure, but the conversion software must have an instant control function, which signals if a record is corrupted.

From time to time, it is worth checking the records already migrated, and as we said above, in the long term we must migrate again the records because the IT environment changes constantly. Migration must be done in a systematic way, for instance periodically or before a significant IT change (e.g. before changing the operating system), and not in a campaign-like manner when we face the risk that we cannot completely convert the obsolete formats. The best solution is when we choose long lasting formats, which do not obsolete every 10 years.

Emulation

Emulation means that we imitate the original IT environment (hardware, software, operating system and the necessary applications) by an emulator (software made for this purpose), thereby we can able to run the original application software (and read the old e-records as they were) in our new IT environment. Emulation needs high IT skills because we must know well the original IT environment, and we have to describe the so-called emulator specifications, i.e. the attributes of the original computing platform (speed, display attributes, tools and peripherals, etc.) which will make possible to programme the current emulator in the future, too[6]. By emulation - as the adherents say, who are usually IT experts - we can eliminate the imperfection of the strategy of migration: the form of the records may considerably alter, authenticity and usability may be doubtful, cumulative data loss may occur in case of consecutive migration.

All these are true; however, archives are reluctant to apply emulation strategy. Archivists do not think that the migration deficiencies listed above would be so serious, this is why they define the record's essence; they say that authenticity is not ensured by unchanged form or bit series, and consecutive migration can be avoided. On the other hand, the complete usability of the emulated records is not an unambiguous benefit because there are many functions (e.g. entering new data, amending, erasing data, and queries relating to workflow) in the electronic records management systems of the agencies, which are irrelevant or undesirable in the archives. According to David Bearman, we have to preserve the records and not the functions of the IT system, and we should not mix information up with the record[7].

6. You can find a good summary about emulation in (Rothenberg, 1999), although he is not objective at all.

7. Quoted by (Granger, 2000).

Emulation has other disadvantages, too:

- To be familiar with the increasing number of application software makes more and more difficulties for both archivists and users. Several hundred applications can be accumulated in an archival institution in a few decades.
- As IT environment changes constantly, we have to create more and more emulators, which must be renewed from time to time. (This can be considered as a kind of migration.)
- Applications are often legally protected and involve licence fee, which, in case of accumulation, can be expensive.
- Records managed by original software mean separate data sets, which prevents from or encumbers establishing a consistent standardized registering and retrieval system as well as access through Internet.

Nonetheless, we cannot completely reject using emulation in archives because some e-records may have such important functions, which can only kept by emulation.

7 The new archival science in the e-records' era

The digital world has transformed archival science because as we saw above the records have transformed themselves: their structure has become more complex, their constituent elements have separated and their importance has changed, metadata have got a highlighted role, the technical aspects of the records have come to the front, and - I have not mentioned yet - standardization has become a basic requirement.

The new archival science, logically, has adopted a lot from information science, and this is manifested in the fact that nowadays standards (ISADg, EAD, OAIS, etc.) have become the new milestones of the new archival science. Standards have prominent role in IT because the capability of the computer to abstract is extremely limited - although it is developing fast -, and even a little deviation from the (programmed) standard results immediately in error. This is why software do not accept "extraneous" formats; this is why "extraneous" programmes do not cooperate. The solution is standardization.

I mentioned earlier the difficulties that the loss of the third dimension and the explosion of the quantity of the records cause. These two factors have deeply affected the archival methods.

Records displayed on the computer screen appear in two dimensions, and this makes considerably difficult orientation. If we find several thousand unordered record files in a folder of our computer, then we can consider them as lost. If we face this quantity of unordered records in paper format (a few boxes), arranging them is not hopeless at all. We can keep order and orientate in, search e-records if each of them are provided with metadata, what is not an unconditional requirement in the analogue world.

At the beginning archivists endeavoured to describe every piece of document. The description - that nowadays we can call making metadata - served both access and registering: when they provided the records with identifier and described them, this fact proved their existence and fixed their position in the collection (authenticity), and made them findable. Such typical medieval descriptions (finding aids) were medieval and early modern catalogues and descriptive lists. The rules of scientific examination of single documents was defined by diplomatics.

The development of literacy resulted in the fast growing quantity of records, which needed to elaborate new systems of registering and retrieval. The more and more increasing ecclesiastical and state bureaucracy had gradually left the description of each record, and switched over describing record groups arranged in subject, chronological or formal order - they were called series -, and later it "invented" the case file, the group of documents relating to a specific action or person. In the 19th century, when state administration expanded more and more and the quantity of the records increased even faster, the above-mentioned artificial series were no longer suitable to retrieve information. This is why archivists developed the theory (and the practice) of the principle of provenance, which enabled users to find safely relevant information in a volume of records, and by which the authenticity could also be ensured.

Archivist, who follows the principle of provenance, states synthetic information deriving from the outer structure, the content and the context of the records, and based on this makes evaluation and description. The origin is always the fonds, a product of a record creator, whose business, legal status; function, etc. are the mentioned synthetic information. Based on the principle of provenance archival analysis is done from the fonds downwards, i.e. from the general to the specific, always keeping in mind the connections of the records (aggregates), their intellectual unity. This concept and method was clearly expounded by ISADg, the descriptive standard of ICA. That is archival science, unlike diplomatics, gives preference to analysis of record aggregates and not to the single records.

We must say that there was a certain "restoration" in the archivists' approach in the middle of the 20th century when, in order to get good-quality records in the future, they started establishing a close relation with the records managers of the agencies. The records managers deal a lot with single records, and their horizon extends only to the fonds. After transferring the records to the archive one of the most important archivist's task is to insert them into the archive's comprehensive record system built up from many fonds, therefore for him the outer structure of the records is more important than the inner one.

However, with the e-records the situation has changed. The radical changes exposed above such as the modification of the structure of the record, the separation and the vulnerability constituent elements, the fact that the corruption is often hidden, the appearance of technical (IT) specification highlighted again the inner structure, the analysis of single records. In the same time since the explosion of the quantity of the e-records[8] made absolutely unrealistic the manual analysis or checking of single records, archives needed such methods (standards) and tools (software) which made possible to provide the records with metadata and to check their integrity an automatic way. This is particularly important after transfer or conversion in order to check data loss[9].

The so far methods and techniques were not suitable for this; therefore, archivists resorted partly to diplomatics, and even more to the information science. The greatest "yield" of the cooperation with the information science was that the data management and procedures have become automatable. The impact of diplomatics is evident in the InterPARES project, it stresses this point[10], and OAIS, a standard for e-records long-term preservation and a product of the information science, has nowadays become a "quasi-compulsory" standard in the archives.

However, to build up an archival system consisting of many fonds needs such a synthesizing task, which is beyond the competence of diplomatics or information science. When for instance an archivist studies the functioning of an agency through its material, parallels it to the political and economic conditions of the time, and, based on this, integrates the fonds into the macrostructure of the archival system, he/she makes such a synthesizing work. He/she makes a similar work when he/she describes this information into the history of the agency, provides the future user with important (and subjective, it is true) information. In the near future, IT will not be able to make such a synthesizing work because it is inflexible and unable to abstract in a way as the creative human being does. On the other hand, it is able to make a different kind of synthesizing task, for example to filter information from an ocean of data or highlight programmed contexts.

In the world of e-records archivist must take into account record as evidence of business and procedure of the agency, but in the same time he/she should not lose sight of his/her evaluating and describing (synthesizing) task. If we cannot keep this balance, either the evidential or the memory function of the records will be corrupted[11].

8. Concerning the estimated number of data created worldwide see for example (Planets, p. 3).

9. In the analogue world it was not necessary, only the outer structure of the records was important, keeping the form or the inner structure was evident if the record remained.

10. See for example (Interpares 3).

11. Abouth the conflict of archival and records management paradigm see (Greene, 2002)

Reference list

Cook, Terry (1996). Archives in the Post-Custodial World; Interaction of Archival Theory and Practice since the Publication of the Dutch Manual in 1898. In: *Archivum XLIII*

Day, Michael (2005). Instalment on „Metadata". In: *DCC Digital Curation Manual 2005*. Available at http://www.dcc.ac.uk/sites/default/files/documents/resource/curation-manual/chapters/metadata/metadata.pdf (accessed on 27.04.2015)

Duranti, Luciana & Thibodeau, Kenneth. *Interpares 2: Experimental, Interactive and Dynamic Records, Appendix 2*. Available at http://www.interpares.org/ip2/display_file.cfm?doc=ip2_book_appendix_02.pdf (accessed on 27.04.2015)

Eastwood, Terry (2002). The Appraisal of Electronic Records: What is New? In: *Comma* 2002/1-2

Granger, Stewart (2000). Emulation as a Digital Preservation Strategy. In: *D-Lib Magazine*, October 2000, vol. 6. No. 10. Available at http://www.dlib.org/dlib/october00/granger/10granger.html (accessed on 27.04.2015)

Greene, Mark A. (2002). *The Power of Meaning: The Archival Mission in the Postmodern Age.* In: *American Archivist*, vol. 65, No. 1, 2002

Heslop, Helen & Davis, Simon & Wilson, Andrew (2002). *An Approach to the Preservation of Digital Records,* National Archives of Australia, December 2002. Available at http://www.naa.gov.au/Images/an-approach-green-paper_tcm16-47161.pdf (accessed on 27.04.2015)

(InterPARES 1) *Requirements for Assessing and Maintaining the Authenticity of Electronic Records, Appendix 2*. Available at http://www.interpares.org/display_file.cfm?doc=ip1_template_for_analysis.pdf (accessed on 27.04.2015)

(InterPARES 2) *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*. InterPARES, March 2002. Available at http://www.interpares.org/display_file.cfm?doc=ip1_authenticity_requirements.pdf (accessed on 27.04.2015)

(InterPARES 3) Available at http://www.interpares.org/display_file.cfm?doc=ip1_atf_report.pdf (accessed on 27.04.2015)

(National Archives of Australia) Available at http://www.naa.gov.au/Images/an-approach-green-paper_tcm16-47161.pdf (accessed on 27.04.2015)

(Planets) Available at http://www.planets-project.eu/docs/comms/PLANETS_BROCHURE.pdf (accessed on 27.04.2015)

Rothenberg, Jeff (1999). Avoiding Technological Quicksand: Finding a Viable Foundation for Digital Preservation. In: *Council on Library and Information Resources*, Washington 1999. Available at http://www.clir.org/pubs/reports/rothenberg/contents.html (accessed on 27.04.2015)

SUMMARY

E-records are the products of information society: many of them (e-mails, blogs) look like oral communication, volatility and dynamics are their characteristics. In e-records, private and public spheres are often mixed which may challenge records managers or archivists. The basic dilemma is that e-records are dependent upon hardware and software which change rapidly, therefore we are unable to preserve them unchanged for long time. IT has overwritten many concepts and rules of archival science. Every alteration originates from four facts: 1) the e-record has no physical aspect; 2) its structure complicated and strongly depends on the technology; 3) it is easy to create, alter, multiply, combine and disseminate them; 4) their quantity increases rapidly. IT has changed the records: they have been dematerialized, their consisting elements (support, form, content, identifier, inner and outer structure, context) have transformed, separated from each other, and have been completed by technical attributes, their role has increased or decreased. The long term preservation is only reasonable if the e-records' integrity (completeness and coherence) is kept what is very difficult because of the above-mentioned reasons (separation of the elements, etc.) and the inevitable conversions, sometimes it is impossible (e.g. web pages with many links). Some archives say that we only need to keep the record's essence. For similar reasons and because records can be altered without trace it is difficult to preserve their authenticity. International research teams (InterPARES) elaborated the requirements and the procedures necessary to the authenticity. In the last decades, archives tried several preservation strategies, such as technology preservation or the so-called post-custodial policy, which did not prove to be good. It seems that the latter has reincarnated in some respects in form of cloud computing. Migration is the most used strategy: the archive regularly converts the files to standards-based new formats. Obviously, this policy has its disadvantages, too, especially when the migrations are consecutive which may result, in long term, in accumulated data loss. Emulation is another proven strategy: the archive

imitates the original e-environment (hardware and software) by an emulator (software); therefore, e-records can be entirely preserved and used. However, the archives are reluctant to apply this strategy because it has a number of disadvantages. In the digital world archival science, adopted a lot from information science, which caused that standards (e.g. ISADg, EAD, OAIS) have become the new milestones of archival science. The greatest "yield" of the cooperation with the information science was that the data management and procedures are automatable. However, archivists must keep their classical synthetic work in digital world, too: to know and describe the record creator agency, its economic, political and societal environment, to build the system of fonds, etc.

Typology: 1.02 Review Article

Submitting date: 07.03.2015

Acceptance date: 09.04.2015