FI MU

Faculty of Informatics
Masaryk University Brno

THE LONG TERM DATA STORAGE (Introduction to Relevant Questions)

by

David C. HAJICEK
Ivo STUDENSKY

FI MU Report Series **FIMU-RS-2005-12**

Copyright © 2005, FI MU **July 2005**

**Copyright © 2005, Faculty of Informatics, Masaryk University.
All rights reserved.**

**Reproduction of all or part of this work
is permitted for educational or research use
on condition that this copyright notice is
included in any copy.**

Publications in the FI MU Report Series are in general accessible via WWW:

`http://www.fi.muni.cz/reports/`

Further information can obtained by contacting:

**Faculty of Informatics
Masaryk University
Botanická 68a
602 00 Brno
Czech Republic**

THE LONG TERM DATA STORAGE (Introduction to Relevant Questions)

David C. HAJICEK
Faculty of Informatics, Masaryk's University in Brno
xhajicek@fi.muni.cz

Ivo STUDENSKY
infocount, s. r. o.
studensky@infocount.cz

November 8, 2005

Abstract

Data storage, especially long-term data storage, is one of the biggest IT/ICT topics discussed in this field today. This document deals with basic classification of electronic data, highlights its specifics versus standard documents and contains overview of elementary requirements on its availability, integrity, confidentiality and storage duration, as well as accessibility time. Together with requirements, this document discusses cornerstones applicable for their fulfillment. Besides these basic factors, you can also find here questions relevant for legislative and normative restrictions put on certain types of data. Finally, we discuss present-day approaches for data exchange and storage, and its usability for long-term storage.

Contents

I ARCHIVES **4**

1 Subject of Archiving **5**

2 Archiving Requirements **6**

3 Legislative Restrictions **7**

II ELECTRONIC DATA AND ITS SPECIFICS **9**

III MOTIVATION FOR LONG-TERM STORAGE **13**

4 Division of Data (by Type) **13**

5 What Data Has Long-Lasting Value and How Long Data Can Keep It **14**

IV BASIC STANDPOINTS **15**

6 Long-Term Data Storage **17**

6.1 Printing and Archiving 17

6.2 Standardization 18

6.3 Computer Museums 20

6.4 Migration 21

6.5 Emulation 22

V PLATFORMS FOR DATA STORAGE **28**

7 Separated Solid Media **28**

8 Specialized Hardware Units **29**

9 Distributed Protocols **30**

VI SOME EXISTING PROJECTS 30

VII DATA SECURITY 31

10 Integrity 31

11 Confidentiality 32

12 Availability 32

13 Non-repudiation 33

14 Time Authenticity 33

15 Discussion about Security Requirements 34

16 Security Methods 35

VIII The future 35

IX Miniaturization 35

X Non conventional methods 36

Part I
ARCHIVES

The least surprising, but notable communication of this article is fact that records (i.e. documents, data and subsequent information) are stored in archives. And everyone has probably very similar image of it: An institution designated for storage and issue of documents.
"Document" and "Archive" are general terms used for data stored both in standard and electronic way. In this document, we hereinafter deal with electronic archives though we usually come into contact with data that can't possibly exist in an electronic form. For example, not many connoisseurs would acquire a scanned image of Van Gogh's painting whereas physicians don't actually mind having X-ray pictures stored digitally, though they originally used "printed" one. In contrast, there also exist data originating electronically that have no importance in solid, printed form (such as software).
The common characteristics rests on a fact that documents are all enclosed, whether we speak about standard or electronic archive. Archives must be divided and kept in compliance with some system in order for us to obtain information from them. Searching through standard archives is very difficult. Therefore, the availability representing one of the basic requirements becomes very questionable and low. In spite of it, the standard archives show several fundamental dissimilarities against electronic archives. This document deals with the most important of them. In order to properly understand questions arising from long-term electronic data storage, we must take these differences into account and treat electronic archives accordingly.

Most of those informed understand weak aspects of standard archives very well. It is noteworthy that archives also offer relatively fundamental advantages in satisfying basic requirements put on archives. The principal advantage is the long-lasting legibility of documents stored in standard archives. It is virtually one of the basic motivation factors for establishing discipline dealing with long-term electronic data storage. Naturally, the grammar and language of the document develop. However, compared with dynamics of informatics it is presumptive that in few years there won't be suitable software to open the document rather than we would not be able to understand its content. Apart from other issues, the following chapters discuss also other standpoints with ambitions to solve these relevant questions.

When speaking about requirements put on archives we must know target users the archives are being created for. The basic segmentation may represent archives managed by state administration and archives managed by private sector. State administration usually creates and manages archives based on scope of legislation (see chapter Legislative Restrictions). Resources of the state administration are usually not so constricted and unfortunately not so carefully protected as resources of private sector. Moreover, archives managed by state administration are usually used only by state administration bodies with no direct access for private sector. Of course there are exceptions, however, the overwhelming majority of private sector representatives must keep documents independently and primarily using own efforts.

Together with state and private division, there is also public and non-public segmentation. This segmentation is more important, because public archives may represent relatively large systems that must meet considerable demands from view of availability, integrity and confidentiality of the data. Likewise, such system's searching logics is usually comparatively more complex. Non-public archive itself may on the other hand be subject to certain confidential degree or store classified information.

Besides an act dealing with classified information (when talking about the Czech Republic), archives may also be subject to other legislative regulations, such as an act dealing with personal information, access to public information, etc. You can find more detailed listing of possible (basic) legislative restrictions in chapter Legislative Restrictions.

1 Subject of Archiving

A perfect archive should not depend on type of documents it stores. It should consider these documents as general data and treat it in this way [1]. Many trustworthy studies set themselves to solve this issue; however, outcome of such attempts is so far unsuccessful. There is no ideal world and similarly, this idea is utopia for the present.

Let's lay the segmentation by format aside for now (we will discuss this later). We will probably want to fill archives with data, for which aging represents evaluation rather than devaluation. It is meaningful to archive documents with time-dependent value only if such archiving is required by law or other obligatory rules (such as standards, company rules, etc.)

However, valuable documents acquire their value over time only if it is possible to prove

time of creation and origin, which is not effortless in case of electronic data.
Situation with data of time-independent value is somewhat easier. These can include for example mathematical proof, chemical formula or technological procedure. These actually do not need to be anchored in time.

2 Archiving Requirements

Archiving is not an autotelic activity. It must satisfy demands of users whose data are stored in archives. Users would want the data stored in archive not to be modified, get lost, stolen (or misused) and would require the data to be available to their owners or authorized individuals at the right time. This results in requirements on search features. An archive should also be capable of storing large amount of data. Thus, the data should be compressed.
It is useful to divide data by their format for subsequent processing and archiving. Multimedia image recordings (such as videoconference meetings), text documents and structured data will all be processed differently. For example video recordings will be indexed and split by indexes in order to store individual sequences to different geographical locations. After querying for the data file as a whole, this data file can then be delivered much faster. If only a part of it is required (for example item 5 of a meeting from 8th of August 2003), the querying users won't need to receive the whole recording and search through it heavily. Plain text documents can be scanned for example using regular expressions. On the other hand, processing and application of searching algorithms on structured data is usually very simple and fast. This means that there will be applied almost no operations before storing the data into archive.
Protocols used for delivery of the data to recipient can also be chosen based on data format. Compression algorithms for image, text, multimedia or other data also differ.
Regardless of type, the data will most probably include metadata, which represent additional redundant data that carry structured information of the data itself.
Together with searching criteria, users placing data queries usually define a time interval on how long is the data important for them. For example medical records must be available immediately at the time of patient's examination. In reverse, a reader may wait for literary work texts even several days without influencing its importance. Also, omission of patient's allergy produced by a certain type of medication may cause harm instead of providing aid. Thus, storage period will be an important criterion. For med-

ical records, (despite a unified time horizon set up for all patients' records storage) it will be necessary to have the data available throughout the patient's life. This time may significantly differ for individual patients. However, after decease the patient's medical records will become just a statistic data. These records may only have a key importance for determination of possible hereditary diseases in following generations. On the other hand, for example land registry records conform to given period, over which is the administrator obliged to keep it. In most cases, there will be no need to extend this specified period any further.
This takes us to another aspect, which puts requirements and restrictions on the subject - legislation.

3 Legislative Restrictions

Let's focus primarily on legislative issues in the Czech Republic. Although the Czech Republic is one of the EU member states, mainly Czech law applies to it. It is possible to divide individual legislative documents by their force to documents regulating availability, integrity and confidentiality of stored data. At present, the most significant document that guarantees personal privacy is

- Convention for the Protection of Human Rights and Fundamental Freedoms (convention).

Article 8 of this convention constitutes everybody's right for respecting his/her private life, home and correspondence. Another document of similar nature is

- Charter of Fundamental Rights and Freedoms (charter; Act 2/1993 Coll., Constitutional Act 23/1991 Coll., which introduces the constitutional form of the Charter of Fundamental Rights and Freedoms, as amended), section 13 of which guarantees postal secret and secret of other documents and records.

This charter surely continues in

- Act 101/2000 Coll., Privacy Protection Act.

Besides charters above, this Act deals solely with personal data protection, determines definition of confidential information and poses requirements on processors of the information. Council of EU solves these issues primarily in Convention 108 from

1981. Another notable document is Directive 95/46/EC of the European Parliament and of the Council (Directive). This Directive defines protection of individuals with regard to the processing of personal data. Unfortunately, legislation of majority of member countries is not in accordance with this Directive.
However, recommendations of EU institutions are not obligatory for member states. In case of Directive 95/46/EC, only about a half of member states legislatives corresponds to it.
Highly specialized, but no less important area in view of archiving is medical care. Although that medical care is very important and recently greatly discussed area, it still poses most of all questions. One of the basic documents that adapts issues related to medical data is

- Act 256/1992 Coll. on Protection of the Personal Data in Information Systems. In 1997, Council of the European Union moreover accepted Recommendation (97)5 for protection of data in medical care, which supersedes previous documents defining operation of automated medical data banks.

Significance of this recommendation is disreputable. However, this recommendation contains notable chapter related to usage of medical data for research and scientific purposes. This chapter determines that data must be anonymous, which makes significant difference between hospital archives and standard archives. Paragraph 3 of article 8 of this document specifies, that restrictions on anonymousness do not apply, if the data is processed for prevention, treatment, diagnostics and other similar purposes. However, only individuals statutory liable of keeping confidentiality are authorized to make these operations. Under local conditions and present situation, it is more likely that archives will rather specialize.
When speaking about statutory determined secrecy, it is also necessary to mention another important act that imposes the strictest requirements on archives:

- Act 148/1998 Coll., on Classified Facts Protection, in the Czech Republic. It is also necessary to mention EU legislative and directive of NATO, member of which the Czech Republic is. In scope of classified facts, it is almost impossible to create an archive that would comply with requirements of the Czech Republic, EU and NATO. For this reasons, we will dedicate only marginal consideration to classified facts along with our effort.

We shall also mention the following acts:

- Act 499/2004 Coll., on Archiving and Archival Service;
- Act 358/1992 Coll., on Notaries and Notary Acts (notary rules);
- Act of the Czech National Council 20/1987 Coll., on Protection of Monument Fund;
- Act 227/2000 Coll., on Electronic Signature, which complies with EU legislation, after amendatory act;
- Act 106/1999 Coll., on Access to Public Information; and
- Act 121/2000 Coll., Copyright Act.

Another notable aspect is legal liability of banks to keep information systems log files. In view of its nature, log can be accessed in the same way as standard text record. However, compared to its volume, it is difficult to satisfy another requirement, which is evaluation [2].

We shall also not miss out a fact that telecommunication services providers are obliged to present these log files to state administration bodies (primarily police authorities), although that providers are not obliged to keep these log files. Also notable is that the telecommunication provider must perform these activities on his own expenses. Police authorities are entitled for these files without juridical permission in extent necessary for fulfillment of particular task, based on section 47 of the Act on Police. On the other hand, clause 7 of article 84 of the Act 151/2000 Coll., on Telecommunication imposes responsibility on telecommunication services providers to make accompanying data anonymous or delete it, upon two months from connection termination. We must remind that law considers this data as telecommunication secret.

However, we don't discuss meaningfulness of law, but possibilities of long-term data storage. Suffice it to say that there is sure enough log data and the longer we shall be able to keep it the better.

The above listing of legislative rules represents only basic requirements and restrictions, which may significantly differ for individual areas and specifics, and frequently even be in conflict. Let's sink deeper into questions and specifics of electronic data and issues arising from its nature.

Part II

ELECTRONIC DATA AND ITS SPECIFICS

First of all, we shall understand that the conception of electronic data apparently contains chaotic series of one and zero digits. And because the carrier is not usually destroyed after being spilled by coffee or dropped out the window or permanently lost in trash on the street, we may come to conclusion that data is indefectible. However, if you damage paper document value of information included in this document may not be influenced in any way. On the other hand, damaged carrier of electronic data, though only partially, means that we permanently lose the possibility to retrieve back all data it stores. This specious data carrier resistance strengthen by hard drive space representing hundreds of gigabytes in user's new bought computer brings us to another assumption that electronic data is readily storable in long-run.

But it would not be a subject field if we could agree with both assumptions. The only fact we can't disprove is that there is no tangible appearance, which sets electronic data somewhat (considerably) apart from standard documents. The fact that electronic data is independent on media and is not solidly attached to it makes it "immortal". In contrast, data is not readable without media so we may even discuss its existence and nonexistence.

The media that contains data can be modified several times without influencing validity, quality of the data or the modification to be tracked down in it. The only way to recognize its migration is an auditing record of individual processing systems.

In addition, lifetime of physical media for electronic data storage is limited. In comparison to standard documents, viewer of data stored electronically may enjoy far fewer times of reading it. The electronic data carrier ages much faster than paper, even if it is not accessed. It is highly presumable that we won't find the data on the media after some time despite our attempts to protect it against humidity, excessive heat, sunlight, electromagnetic waves and other adverse impacts. Howbeit, not so many impacts of such nature act on paper (parchment or clay tablet).

Discussion about value of documents versus their form is very interesting. In case of electronic data, form is usually not determined by carrier of the data. If we overlook

the value (or price) of the carrier itself we learn that in digital world data is valuable only from view of content, not so much from view of its origin or originality. What price would Czech historic documents, such as Chronica boemorum or translation of the Holy Bible from 1613 called Bible kralicka, probably have in PDF format? It is obvious that the price would not come to value of archival documents, but there is still a question if it would not rather represent just a snatch of original masterpiece's value. However, we may assume that for written documents the price is primarily determined by their content. What would happen to value of Mona Lisa if Leonardo created it in Photoshop?

Moreover, thanks to rapid development in IT technological obsolescence of physical media rapidly accelerates and so media become useless. At present, there are no computers allowing the user to insert 8 inch floppy disc or MFM hard disc drive. In addition, there are even no drivers available for old devices. We may even encounter devices, drivers of which are no more supported neither by their manufacturers (usually nonexisting anymore) nor in standard installations of present operating systems despite they were present in previous releases. Besides that this makes data inaccessible, the value of media considerably degrades. Turn to this trend of decreasing value is only possible after maybe several decades when the media becomes attractive and valuable for museum purposes.

Data legibility is also limited during time, not only due to media obsolescence and language and context intelligibility, but primarily due to its format. We often fruitlessly search for software the data was created with similarly as there are no readers or buses enabling connection of the carrier.

Though that digital signature or water stamp may guarantee data origin or an author's identity under certain circumstances, determination of originality represents requirement, which is in contradiction with its principle. Theoretically, we could prevent creation of counterfeits by an author's statement on valid time stamp of an original, however, what would collectors this knowledge of authenticity have for if anyone could own an original masterpiece? An implicit characteristics of data is a fact it can be duplicated and copied - easily and virtually without any possibilities of control, subsequent expenses and extensive effort.

Although zero and one digits don't fade away and are intangible, it is logical and legitimate that its originators, owners and processors try to ensure its uniqueness and prevent its modification. Electronic data are subject to virtually the same requirements

as standard data. Data confidentiality, authentication of origin and time authenticity was solved by ingenious and simple idea of cryptography.
Solution of issues related to long-term storage of electronic data requires equally big idea as cryptography, especially asymmetric, which became revolutionary idea that brought at least partial principals and order of standard documents and texts into "chaotic" background of electronic data.
From philosophic point of view, we must ask a question deconstructing meaningfulness of the idea of long-term data storage. Let's express a wild assumption that things that don't fade away can't be stored for extended periods of time. And what brings us to this thought? Let's take a look at the following example:
Whereas in Ancient Egypt embalmers made huge efforts using alterations to guarantee aging dynamics approaching zero on their client, aging of electronic data (when speaking about physical aging) is straight equal to zero. From principle of natural evolution, genesis and termination, the physical existence of actions and objects is determined by their existence in time. This existence in time can usually be almost exactly defined. For electronic data, it is possible to determine time of its creation with certain accuracy; however, we have absolutely no idea about time of its possible extinction. Therefore, this piece of knowledge might take us to conclusion that things that don't age don't practically exist.
However, let's believe this speculation is wrong and our efforts are at least as meaningful as attempts of mathematicians, who deal with formulation of reality and its principles in algebraic units. Mathematics as source of philosophy tends to express actions formally. Let's indicate advance of mathematician's ideas: Initially, natural numbers and addition operator were sufficient. But after definition of inverse subtract operator, the mathematician came to conclusion that only natural numbers won't suit the purpose and enriched the system with zero and negative numbers. After some time, with introduction of multiplying operator, it was also necessary to divide numbers. This brought about a need for real numbers. Throughout the times, other objects, groups or areas developed together with respective rules and limitations. And because the ongoing advancement has not stopped yet, we may assume that the revolutionary idea of mathematics is still to come and thus it is still important to continue with attempts to imprint the multidimensional reality into imaginary clay tablet of mathematics demarked by X and Y axis.

Let's consider this conclusion as a sufficient justification and motivation for our continuous efforts in a virtual world of electronic documents created so far.

Part III
MOTIVATION FOR LONG-TERM STORAGE

The first motivation factor is constantly developing informational society that urges everyone and itself into exploitation of computing technology in virtually all areas of human activity. As an example, we may mention the requirement for fast, simple and cost-effective way of searching through land registries, trade register, address books, medical records, accounting and economic data or acts and contracts that were so far available only in printed form (or paper form to be exact). Also clerical archives are significant in this context, because they are not subject to legislative rules in many countries.

However, documents of this kind are not only processed and searched, but also stored for long-run. And what does this long-run stand for? Let's consider decades as units of measure. We may apply this unit of measure for instance to economic and accounting data and even hundreds of years to contracts (private, commercial, statutory or international) or land registries.

Archived data must be kept and found in the same state the author stored it in: it must be intact, contain the information it stores and be available in reasonable time. Availability and reasonable time represent key concepts related to long-term data storage. The data must also be readable and clear.

4 Division of Data (by Type)

As we will thereinafter demonstrate, the greater platform independence we try to guarantee when creating long-term data stores the more detailed knowledge about each platform and data specifics we must have.

We have already discussed motivation for data division previously. Let's say we have accepted the unpleasant aspect that resides in adding other information - metadata -

and the associated increment of data volume. Now, what categories shall we consider? It is

- Textual data - documents, source code, and configuration files etc. (disregarding various formats and special instances with characteristics of binary files that are not included in this category);
- Structured data - databases (Also not distinguished by particular database technology used for creation. Division by subtypes suits the purpose only for databases of SQL92, OODB standard and OLAP stores);
- Image data (Disregarding image format. It is necessary to take into account also other characteristics, such as embed water stamp, that prevent from using lossy compression.);
- Binary data (SW);
- Multimedia data (audio, video).

In many instances, it is useful to solve issues related to type of data. For instance electronic registries of state administration bodies must solve relatively remarkable issues with announced competitions of tenders, because they are obliged to accept proposals in electronic form. Complications arise when a bidder presents proposal that contains viruses. Infected files may not only cause problems at the time of take over and reading, but also at subsequent archiving - however, only if it's necessary to store such type of data in long-run. This takes us to another question associated with the subject:

5 What Data Has Long-Lasting Value and How Long Data Can Keep It

First of all, we should distinguish between value and price. Often, it is quite difficult to express price of data. Primarily in case of an organization key data or data, preservation of which is required by law. Organizations may apprize its data for instance by lost profit or market price of competitive advantage. For the latter group, the price may be expressed for instance by determined recourse. Therefore, it is more appropriate to speak about value, which is usually predetermined within given scale. It is pointless to list data with no value. Let's rather focus on data, which represent subject of our

interest.
The value will definitely be preserved on data in registries, land registries, classified facts, notary records, common commercial and noncommercial contracts, international contracts and acts. Also data originating, existing and possibly expiring in electronic form will certainly not lose its value. This may include for instance work of authors who don't publish in standard way. The value will surely be also preserved on other literary work that gets digitalized only after its creation. When speaking about copyright law, we should mention that data representing historical relics will not lose its value, though it is still not so usual to encounter it in an electronic form.
It is indeed very difficult to determine the value in this case. Since it is not about value of an individual instance of data, but the value of information we can acquire from it, though even this value will differ for individual user groups.
We may possibly determine the value extent of technological process used for production of hamburgers at McDonald's or nuclear bomb. But how about above mentioned relic that represents national historic heritage?
We shall surely agree on fact that determination of data value should be result of some long-term analysis rather than fly by the seat of pants.

Part IV

BASIC STANDPOINTS

For purposes of long-term storage, it is useful to mention terms data exchange and its short-term storage. What attributes can we find with this activity? It is usually associated with exchange of large data volumes, which can't be transmitted using standard means, such as e-mail. The technically simplest, but the least comfortable and effective way is exchange on removable media. Since capacity of some media may be limited we may encounter situation where such a transmission would not possible.
Users in their environment often simply share their documents or parts of hard drives. But this is one of the least secure ways of data exchange. Another possible way to exchange data is use of individual storage spaces, such as FTP archives. However, these archives are closed and their capacity is also limited. This demerit causes the data to be redundantly stored to several locations at the same time, which may be quite trying for users in view of the limited capacity.

Data closed in individual applications are often streamed into OLAP stores that enable to share data with other users within organization. However, from view of global archiving the data is closed the same way as in applications it is created with.

The intention is to build distributed structures that would be as much open as possible. One of such projects is for instance DataGrid (see http://www.ten.cz/doc/2003/zprava/ for other projects developed under patronage of Cesnet within the 5th general EU program). Everyone is probably familiar with the grid projects that make use of unused capacity of user PCs in calculations. Participation in such projects is voluntary and calculations are performed at the time the processor is not used by applications launched by its owner. Idea of such sharing is relatively interesting and powerful and even usable in field of data storage to certain extent.

After placing request for data storage, the capacity of participating stations can be used for data indexing or compression, hash counting, partial data cryptography and accompanying calculations. On data search request, stations may be of significant help with searching and presentation of the data (see Emulation further in this document). It is not surprising that reader of this document could come up with the following logical idea, though inapplicable in field of long-term storage: Idea to use also hard drive space besides processing capacity and short-term RAM memory. Similar idea recently originated in BetterFind project, which was developed on "Druzba" college of Masaryk University in Brno. Group of students has used college network to create system that is accessible from a web portal through a thin client. This client enables access to every college resident, who runs an FTP service on his or her computer and sets up standardized user account protected by password. In addition, residents are required to update information about the data stored on their FTP servers using specified tool and at regular time intervals. Regrettably, we can straightway point out the biggest weakness of such solution: If they fail to observe these conditions or server drop-out occurs, users would lose data stored on the server. Though, this system is not meant to become long-term data storage, but only an environment for easy way of data sharing and exchange, soon or later, solution designers making use of this basis would have to face this issue.

The above problem can be resolved for instance by creation of system with several redundant data storage sites that would extend occurrences of data instances in case that count of data instances drops under given threshold. It is unlikely that all servers

containing the same data would encounter drop-down at the same time. However, the spatial complexity of such solution and computing demands related to such redundant system appear disproportionately high.
Another option is to build standard distributed closed architecture for instance on Storage Area Networks basis. However, this solution would also grant access only to enclosed group of users, which is undesirable. This would only do the job in large and powerful institutions and organizations and represents very expensive solution, though widespread at present.

6 Long-Term Data Storage

We have already touched questions related to data storage few lines above, now let's take a closer look at possible ways to solve our subject. Although electronic data is not identified and tightly bound up to media it is found on, it can neither exist nor be interpreted without media. Therefore, the first available standpoint would be explicit binding of data to media.

6.1 Printing and Archiving

The easiest way to bind data to media is to print them on a paper. After printing, data is archived - that is, in a standard archive. Persistence of this way stored data is guaranteed for hundreds of years, unless disaster occurs that would damage it or its parts.
Printing and archiving takes us back to complicated archives organization structure and catalogue cards. Back to long waiting intervals for required document, lost documents, damaged parts of archives and hours spent in reading rooms.
Printed documents also lose their characteristics - such as interactivity, multimedia features and dynamic aspects. Nonlinear documents, such as hypertext, become completely useless after being printed. Flash presentation becomes just a pointless image and encyclopedia courses just a pile of paper. So it is virtually impossible to print it.
To certain extent, we also lose irrevocability and trustworthiness. Due to inability to print control characters, we for instance lose an option to validate digital signatures. In this context, we must be aware that not in all countries the legislative rules allow to attach digital (electronic) signature only to documents viewable by the signer. On the other hand, this may strengthen integrity. Printed documents can't neither be changed nor as easily counterfeited as electronic documents. Besides this, time credibility in-

creases as well. In real world, time can be examined based on other attributes than in electronic world, such as chemical analysis etc.
Therefore, printing is actually not the solution.

6.2 Standardization

Another standpoint arising from long-term data storage issues is to archive just documents that meet certain characteristics - that is meet given standards. This limitation can be for instance requirement to encode using UNICODE.
Unicode is a 16-bit character encoding. It is popular primarily because it includes alphabetical characters of many languages and thus represents considerably universal encoding. It is already widespread and it successfully expels other ways of encoding. It is commonly used in communication protocols and even in mobile phones. Here we could refer to many other areas of usage.
At present, the UTF-8 is the most widespread version; however, due to its insufficiency it is presumptive that in the future we shall rather use its successor UTF-16. Unfortunately, even the Unicode encoding family is pretty divaricated and virtually every IT or telecommunication area uses its clone (even mobile phones - UCS2). Despite today's relatively favorable situation, duration of Unicode's increasing trend is uncertain. What lifetime can we expect for Unicode? Fifteen years? What was for instance the lifetime of EBCDIC or ASCII-7?
Some higher grade limitation may be for example requirement to store documents in XML format together with digital signature. SGML is a widespread and flexible format for electronic documents. SGML is a metalanguage designated for textual, structured document description - segmentation. SGML itself does not provide us with a solution for presentation of non-textual data and dynamic nonlinear (interactive) documents. This language became a sort of text processors intermediate language, which significantly facilitates text conversions. Conversion is the biggest concern of IT users today. However, it is improbable that SGML could supersede all other formats in current market.
Storage method that would fulfill as much criteria as possible could positively impact choice of storing, speed of storing, speed of search functions, choice of compression methods, transmission protocols and the system complexity in general. Well, for this particular solution and under particular circumstances, it seems that support in standardization offers the best possible way for creation of contemporary solution, which

could be replaced by truly long-term answer in the future (see efforts of W3C, OASIS organizations). However, we should take into account that there are huge numbers of various standards with really short lifetime.

An example of long-term storage could be an international OpenEvidence project. Besides other companies, also Czech IT companies participate in it (such as PVT). Primarily, this project deals with archiving of simple Unicode encoded and XML structured text files supplemented by digital signature (see [7]). For the record, OpenEvidence project treats the generally proverbial question of electronic signatures validity in the following way. Document is supplemented with archiving time stamp, which extends validity of signature for the validity period of the stamp. Before expiry of the previous time stamp, the document gets new time stamp and the whole process repeats over and over [8], [9]. Unfortunately, this approach increases the number of time stamps that must be maintained, because validity and authenticity of each one of them must be declared when using signature and time data. This makes authentication very demanding from view of space and time.

OpenEvidence project represents purely specialized archive and creators even treat it this way. Issues of long-term storage are, however, of more general nature. If we chose similar approach, we would have to create dedicated archives for each type of data and all combinations of requirements. With all respect to Terry Fox, this method would be as useless as his prospective attempt to beat a race car in speed.

When choosing this kind of approach, it is very important to select standards operative in long-run, if there is any. One member of this small group is for example SQL92 [4]. SQL92 contains just description of the core component. Individual databases differ by their extension functionality. If restricted just to standard, we would lose information and frequently even the necessary functionality. In addition, this standard usually gets enriched by many other approaches (PL/SQL etc.) If we include also OODB in this listing we obtain completely different approach for structured data processing (though we still talk about databases). Let's finish with observation that there is no automatic conversion available.

Many companies offer and stand up for OLAP as a nostrum. Although it represents a promising tool, we should keep in mind that these tools are usually designated for users and their data presentation purposes. Moreover, method of data storage makes these tools also relatively specialized, thus generally unusable.

They represent members of a small family of standards, which are based on mathe-

matical model. Regrettably, the overwhelming majority of standards represent ad-hoc approaches, which are closely bound of particular platforms and technologies.
Very unpleasant is that for instance in field of audio and video formats, text editor formats and typesetting, hypermedia and other formats, the development is so fast that expectation for some final and long-term operative standard appears to be impossible.
Another handicap of this approach is of course significant limitation we put on stored data. By standardization, we disqualify large classes of data types that can't satisfy associated requirements. However, we can't cast the idea of standardization aside - quite the contrary. Support for long-term standards is determinant (or at least very helpful) even for other approaches, especially if metadata is used.

6.3 Computer Museums

This approach considers existence of old data centers. User with intention to review data from certain IT period could choose a device that the document might have been created with at its time of creation. The mandatory prerequisite is also presence of suitable software. And while not all IT users would be capable to determine the device for viewing the data it would have to be supplemented with metadata. Or, there would have to be data classification tool, which could determine technology based on data format and then direct the user to relevant machine. Development of such tool would require usage of large and updated data type databases.
Besides demanding database management, there is also problem arising related to maintenance of the old hardware. Each piece of hardware is limited by its lifetime and repairs might be impossible in the future. It is also assumable that for really hoary machine pieces we would not be able to obtain necessary components. As mentioned above, media lifetime is also limited. Thus, this way of resolution is not much cop either.
Apart from other factors, extended response times and access speeds of old technologies cause low data availability.
It depends whether we want to build computer museums or architecture with truly long-term lifetime that brings attractive added value. We would probably like to achieve a state allowing us to replace every component (whether for failure or obsolescence) without influencing functionality of the whole (at run time) and substitution for fresh technologies and application of new approaches. This means the greatest possible platform independence and stability of the whole. Naturally, stability will be consider-

ably weakened due to data loss we risk with removal of archives or their parts.
Let's take a look at solution possibilities that are closer to reality.

6.4 Migration

Migration means conversion of certain document format to other format at time of its obsolescence. This means conversion into some new format as long as the original document is readable. These new formats the data is being converted to should naturally consist of the smallest possible group and have the longest lifecycle available.
Migration can be divided into three basic categories: Fully automatic, semi-automatic and manual. Automatic migration, as determined by its designation, is performed mechanically without human interaction or control. It is usable and used rather occasionally - for instance when converting specific encoding into Unicode.
Semi-automatic migration is performed mechanically as well; however, its results must be checked manually. Most conversions involve this type of migration. Example of this migration type may include conversion of economic data from Microsoft Word format into XML structured documents.
Manual conversion is the least comfortable mean of data conversion. It is usually performed if mechanical migration is ineffective or impossible. The typical example is conversion of data from SQL relational database into object based OODB database.
Quality of migration process and frequency of its necessary recurrence is influenced by choice of new formats. Regrettably, there are no explicit rules for process of decision making. Moreover, we can't learn from the past either. Each choice brings new issues we must take into consideration as well as new risks. We may observe that individual conversion steps differ in various versions. Different order of these steps in new and previous versions may also make automatic conversion impossible and require us to proceed manually.
If we overcome issues related to migration execution, we can ensure legibility of documents even in future. However, this approach is not ideal either. For instance, we can neither guarantee preservation of information nor perform trivial checks for possible losses. Verification through reverse conversion into original format is usually as difficult as the conversion itself.
If we decide for data migration we must also pay attention not to miss the time the document is still readable. If this occurs, we lose the document forever. Therefore, it is necessary to continually monitor and control all data types and existing documents. For in-

stance, we must migrate documents consisting of several types of subdocuments (such as multipart e-mails) whenever some of the subdocuments become obsolete. Thus, migration becomes never-ending process, though still labour intensive & expensive, with uncertain results and prospects we would happily lose information only after performing several migrations, so, frequently resulting in failure.
All facts mentioned above make migration demanding and extended process with uncertain results. Moreover, manual conversion can put us to another risk represented by possible distortion or enrichment caused by processor's passionate attitude or context. Resources processed this way become further inapplicable for reputable work. For instance, Egon Bondy who engaged half of his life in investigation of Buddhism also based his work on original documents that were not enriched by outside renditions. However, his advantage was that he could use paper documents - something we can't rely upon today.
Added to that, many companies (even Czech financial institutions) that decided for migration of obsolete format documents, driven usually by obsolescence and substitution of current information system, today deal with troubles, pick up mistakes in new documents and search for other documents in standard archives. This process usually takes years, so it is small wonder migration is not construed as a suitable approach for long-term storage.

6.5 Emulation

Emulation is an interesting approach initially discussed in [5]. In brief, emulation is based on creation of virtual computers and imitation of outdated programs and systems that enable to launch and view original data in original version. Important and attractive characteristics of this approach is the accent on original data version that allows to preserve its integrity, irrevocability and digital signature - simply all original properties. Due to possibility to copy data to new media or archiving systems it is easy to check for possible damages using any integrity check algorithm.
Naturally, one of the requirements is to run emulation on new machines (with attributes and I/O devices we are familiar with in advance). This requirement brings necessity to create a general model of the computer and I/O device operative in long-run. The emulators could be then created inside this model. The advantage is that the emulator would be created just once for certain data type. After that, all data of this type would be available.

The limitation may be a need for development of general model interpreter for currently used platforms. Development process of long-term applicable formal document descriptors - the metadata (such as substitute for media labels) [3]) - can also be quite trying and demanding.
We must be aware of emulation possibilities:

- Programs: We should emulate only programs that can read original data. An example may include MS Word emulator for text documents. Non-existence of detail specification for all such programs creates an obstacle. Most of the programs commonly used by commercial entities and statutory organizations are actually not open-source, but commercial programs, producers of which usually don't reveal details about. Cooperation of all producers would be required. There would have to be a lot of effort spent on creation and maintenance of such database. This activity would be nearly impossible without support of autonomy statutory and supranational bodies. Leaving the provision of formal descriptions to voluntary participation of producers would probably also not do the job, thus, we wouldn't do without legislative emendations and subsequent enforcements (naturally, in supranational scope). It is improbable that commercial subjects (producers) would come up with emulators spontaneously, especially if it wouldn't generate profit.
  How large would have to be the working team able to process groundwork for all programs emulation, maintain it and develop and maintain the emulator itself? Some producers are insofar innovative that new emulator instance would have to be developed for each new version of their software. For instance, NASA images sent from Mars were only viewable in Acrobat Reader 2. In later versions, images did not display properly (this is another example of situation when data conversion is extremely inappropriate).

- Operating systems: We talk here about emulation of operating systems that enable us to launch programs designated for reading of and manipulation with original data, such as MS Windows operating system for MS Word.
  Number of operating systems is massively lower against programs, this means less work spent on creation of emulators. This is also ensured by more extensive documentation, which is provided in order to enable development of operating system compatible programs to other producers. The general model interpreter must emulate original input/output device, including its base characteristics, such

as sequential/direct access, speed etc.

Installation of original programs into operating systems is not completely commonplace. Each operating system handles it differently. One of prerequisites arising from this fact is necessity to deploy long-term applicable descriptions or installation instructions. However, these rules and configuration must also be readable in long-run, or can be replaced by automations performing installation and configuration without human interaction. A watchful reader probably feels that problem of this approach will be very much similar to conversion - it would not be always possible to create such automation.

- Platforms: The lowest emulation level is represented by platforms individual operating systems are operated on. Since the number of platforms is lower than number of operating systems, it is possible to operate several operating systems on each one of them. There are several operating systems designated for more platforms, however, all of them actually represent different systems that just appear homogenous.

  Platforms are well documented, relatively uncomplicated, thus, their emulation is usually not difficult, including their basic characteristics. However, installation of original operating systems and programs into these systems is difficult. As in case of operating systems emulation, it is also necessary to have access to installation instructions or perhaps automations that perform installation with no or minimal user interaction.

  In any case, all efforts are useless without support of software and hardware producers (which is usually not an issue): Copyrights for software being installed into emulated platforms, whether locally in clients or centrally on servers, are usually property of their owners, for nonproprietary and frequently even for proprietary solutions, thus, licensing policy is question of mutual agreement between provider and user (or operator of an archive). As mentioned above, software producers will not be always willing to provide data type descriptions or proactively participate on archive development.

  So, what would be the data like in case of unsuccessful implementation of this approach? The first possibility is to keep all necessary information with the data: Data armed with metadata containing formal format description, besides other information, are stored into storage space. This way, the data can be stored in full context - encapsulated with everything necessary for reading (launching). This

package could contain platform emulator, original operating system, original program, original document and descriptive metadata.
With the second approach, only formal description and unique identifier is stored within metadata. The identifier is then used to retrieve everything necessary for reading from central storage. In comparison with the former, this approach saves space, access time and time and costs of the solution. On the other hand, it involves creation and long-term existence of emulator storage.

No matter which of the above approaches we choose, it is evident that long-term storage makes general use of simple data exchange principles and also leans upon present-day short-run storage methods.
It is necessary to add that besides the three above basic principles, which are seldom used in the field (or rather exceptionally); the short-term variation with extended storage time is considered to be long-term storage as well.
However, one of the basic differences between short-term and long-term storage are, besides others, extended access time and strong demands on platform independence. This puts stronger demands on solvers and their knowledge of individual data type specifics and storage approaches. Let's close this chapter with an example of research that dealt with storage of one particular data type - videoconference outputs (VUT Brno, GiTy, a. s., resource [12]):

Example 6.1. *Besides software based user friendly interface created for videoconference system users, the research also involved development of a scheme for recording and long-term storage of videoconference records. For on-line access, the storage period was set to 10 years. Obviously, the research was not about long-term storage at all and furthermore dealt just with a dedicated location.*

You can find the general scheme on the following picture (see Picture 2: General scheme). Primarily, the research dealt with output recording, participation of users with no financially ambitious sets available and indexing of videoconference recording for subsequent file partitioning (see Picture 1: VCF output recording scheme). Purpose of the scheme on the first picture is to introduce context of the whole solution; we will not discuss it in more detail. In the Common Data Storage Environment part, the picture changes into data storage scheme, which is fundamental for us. Projects solving similar questions are also handled by FI MU [13]. The laboratory deals with videoconference output questions in more complex and more detail way, due to larger capacities.

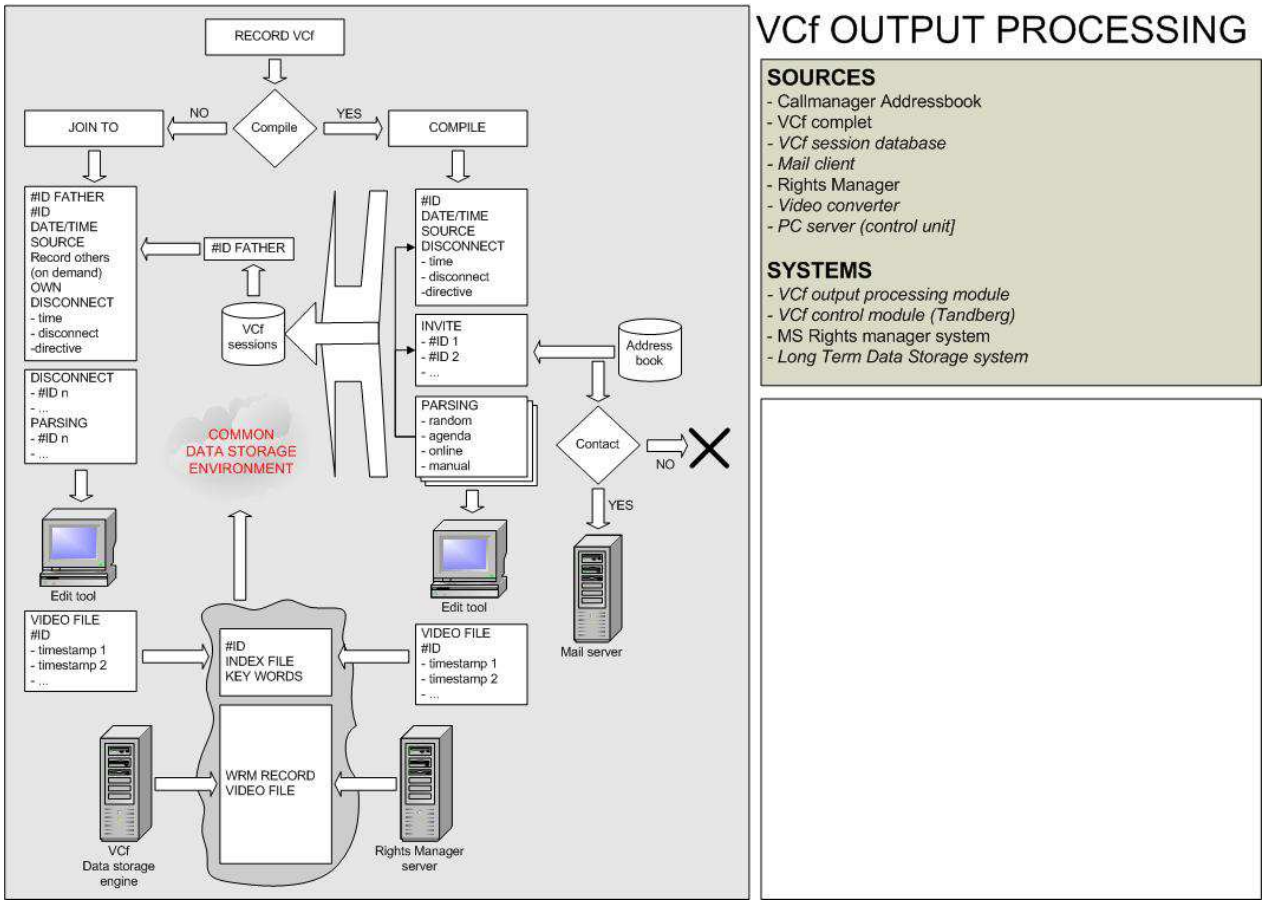

Figure 1: VCF output recording scheme

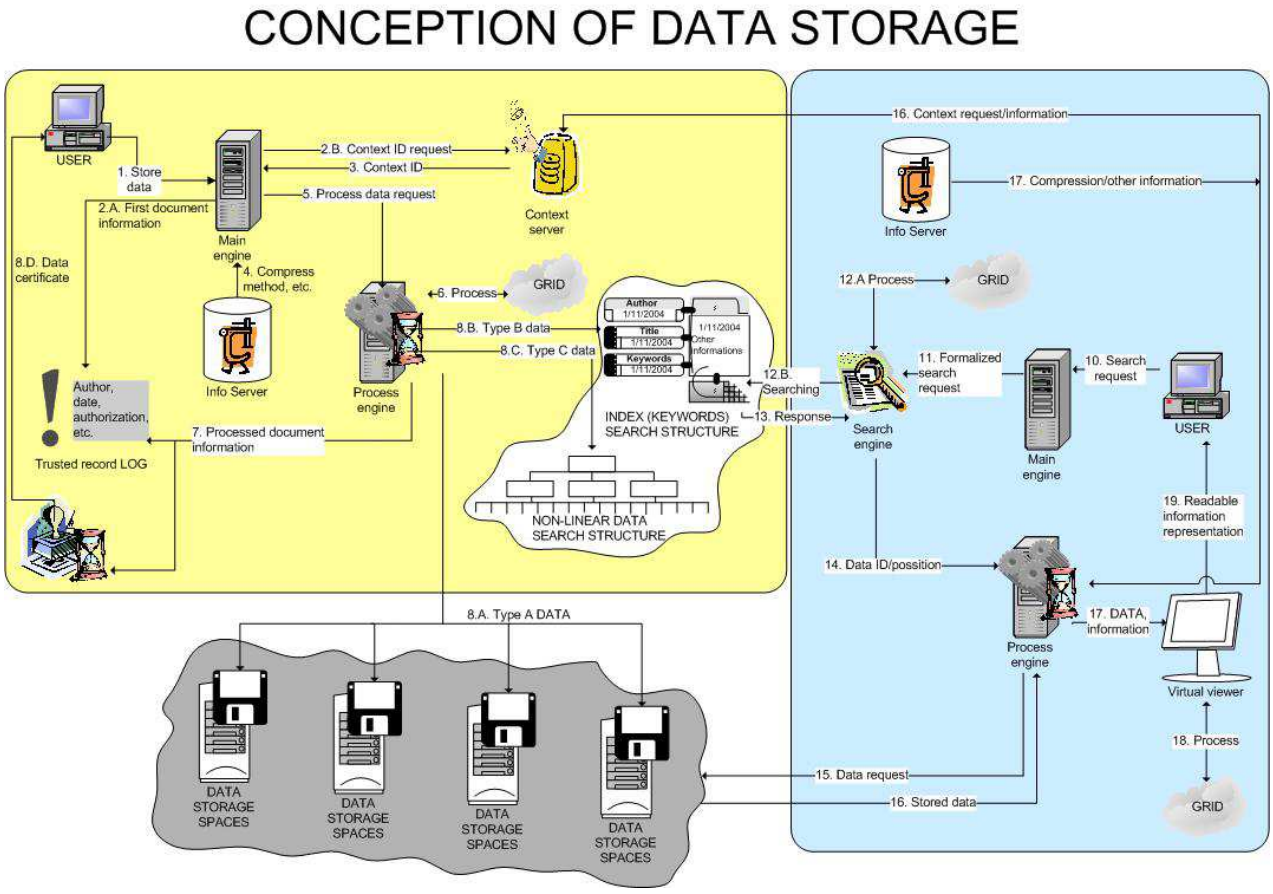

Figure 2: General scheme

As it is apparent from Picture 2, the project makes use of grid calculations. However, storage space is already proprietary and does not rely just upon optionality of participating users. Originally, even this alternative was considered. It included several user groups divided by reliability (i.e. by users who pay for provision of services, or users who are bound by contract to guarantee specified storage space availability), that would influence hardware type or site designated for data storage. However, from view of administration, such scheme was so demanding that developers completely turned away from it.

We may also see that searching is performed not only in database and using regular expressions, but also using non-linear (tree) search structure. The scheme makes use of context server (i.e. data format formal description storage uniquely specified by identifier, which becomes part of metadata upon call for data storage).

Grid calculations are used in phase of storage, data processing (such as compression), subsequent searching and processing to output.

Part V
PLATFORMS FOR DATA STORAGE

Regardless of data storage environment we possibly build, we will always depend on technologies and memory media. Let's mention at least the fundamental types.

7 Separated Solid Media

Usage of separated solid media is one of the most common methods of data storage. CDs, DVDs, discs, tapes, flash memories or hard drives, all of these are probably the best known representatives of this category. Recently, we could also include floppy discs (or punched cards :-) in this category. Separated media prove to be somewhat insecure. If we ignore the possibility of easy loss or theft and resulting demands on media archiving, there is still a threat of damage caused by routine manipulation. There are also considerable demands on constant temperature, low humidity, protection against electromagnetic waves etc.

The media's lifecycle counts among years, thus is considerably limited. The National Media Lab's research showed that media, such as hard drive, tape or CD, which are accessed randomly without an option to plan the access time ahead, have lifecycle of approximately 5 years [6]. The research also demonstrated that their quality degrades regardless of whether it is accessed or not.

There is also an important fact that media frequently contain errors. This results in need to duplicate stored data, ideally altogether with physical storage in separated geographic locations, or at least objects. This prevents damage when natural disaster or similar destructive element occurs.

We should also not forget about necessary migration of stored data onto new types of media before the original media becomes completely obsolete. As we mentioned earlier in this article, migration must be performed at the time the old media is still readable using usual computing means. This obsolescence is question of years. After that, we expose to risk there will be no devices capable of reading or drivers available for current operating systems.

Another observed demerits are extensive access time, archive limited search capabilities and even inability to access data in some cases (due to geographic distance between

archive and requesting entity), difficult manipulation (such as need for careful transportation), small capacity etc.
Although manipulation with such media is difficult, we may still observe this approach in some special cases of larger information systems. The typical example is a system designated for preservation of and manipulation with classified facts (subject to Act 148/1998 Coll.), which considers usage of PC station with removable hard drive as one of the alternatives. When the drive is unused and station is idle, the drive is placed in a strongbox.
Specialized hardware units, which are however also based considerably on separated media, represent one level above the separated media.

8 Specialized Hardware Units

These units represent usually relatively large computing centers, i.e. closed, secure and maintained systems for electronic storage. They make use of various data storage methods - array fields, server farms etc. No wonder these centers are both labour-intensive & expensive, thus pretty demanding in view of administration. Its price usually directly depends on chosen producer of computing center's components. Each upgrade, whether complete (rare) or partial, involves quite large investments; small upgrade=big investment. In comparison with separated solid media, the price is unimaginably bigger.
Naturally, specialized hardware unit's capacity is bigger then capacity of separated media. However, even this capacity is limited, though it can be extended. The demerit is the above mentioned investment, which depletes informatics budget of an organization for extended period. The interesting fact is that investments for X years ahead of time usually satisfy requirements just for period expressed as X/2 years. This means that approximately half of its existence the organization deals with issues related to (long-term) data storage.
Moreover, storage of data in one place, which is usually the case of these systems, causes data vulnerability, lower output of the whole (for instance on complex search or requests simultaneously placed by several users) and unfulfilled demands on system response. Relative to particular issues and conditions, these systems may represent suitable solution.

9 Distributed Protocols

The family of platforms closes with distributed protocols. It comprises of several centers and end points, which are deployed in several sites. The whole system exploits computing force and capacity of several geographic locations. This makes the physical maintenance burdensome, though, most of the configurations can be performed remotely. In spite of it, this system brings many benefits. Besides possibility to integrate multiple organizational units, which cuts down expenses, it also streamlines data duplication and helps to avoid data losses. An important aspect is outside transparency of the system, though, it appears to be a whole.

Insertion and acquirement of documents is relatively easy - it is usually performed through some user friendly interface (recently, also using thin client). Components of data can be stored into various sites, based on needs and its characteristic, working environment etc.

It is the best environment available that has high potential at present. Although distributed data warehouses are relatively common phenomenon, its usage for the real live long-term storage still appears more a as pie in the sky.

Part VI
SOME EXISTING PROJECTS

Let's pay attention to some interesting ongoing projects that relate to large data volumes and its long-term storage. The first place belongs to DiDaS project [14], which develops and works on groves of academe (of FI MU in Brno, besides others). It is a transparent system, which allows a user to easily insert large volumes of data as well as retrieve it back from the system.

Once the document is inserted in the system, the user gets a descriptive file, which specifies location of the document. Acquisition of data from the system is possible only with this file, based on its ownership. When stored, the user may define number of document copies that will occur in the system. Unfortunately, the current version does not offer choice of localities the copies could possibly be stored into. If not restricted by the user, inserted document is automatically divided into several parts, which are randomly stored into separate system nodes.

The DiDaS project is the herald and a step in right direction. Still, this system is designated rather for data exchange than its storage (long-term is absolutely inapplicable) and has no other ambitions for the present. Data is kept just for a given period (14 days, but it is strongly configurable) and the system does not allow searching.
Another FI MU Brno research activity in field of large data volumes manipulation is a project dealing with digitalization of medical documents - MeDiMed [15]. However, it is still just a specialized warehouse, despite obvious ambitions to provide long-term storage even of large data volumes.
One of specialized solution is the above mentioned VUT Brno and GiTy, a. s. research, which is focused on truly effective long-term data storage. However, activities are still in embryo and only partial concepts and suggestions are being practically verified.
And finally, a word about PVT, a. s. activities put into this field. Besides participation in OpenEvidence project, this company directs at building of distributed archive [16].

Part VII
DATA SECURITY

Integrity, confidentiality, availability, irrevocability and time authenticity are all troublous areas that may significantly influence the investigation.

10 Integrity

The basic prerequisite and requirement is a possibility to acquire data in the same form it was stored in. Standard paper archives have this characteristic by default (unless some parts become damaged). The present approaches are based on cryptography algorithms and hash functions derived from it. As mentioned above, this approach is not actually optimal. There are interesting options opening up in algorithms based on quantal calculations, however, at present there is no equipment available we could use without objections. Neither hash quantal algorithms are currently being investigated.
Another approach relies on foundation of independent, trustworthy third party (TTP), which would guarantee data integrity. This is usable in world of "tangible" documents, such as for notary services. However, there is also possibility to blaze a trail of collective confirmation, whether standard or modified (see [11]). One way out of the col-

lective confirmation principles, TTP and cryptography could also be trust in electronic notary services ([10]). However, there is no probability of long-term applicability of this alternative either.

11 Confidentiality

Confidentiality should be guaranteed upon issue and storage of documents into the archive. Though it may look irrational, it is a bit easier to guarantee confidentiality, or at least confidentiality of some components, than integrity. Confidential transmission can be ensured by the administrator, who may handle the long-term data storage system we probably won't do without in the future, based on new facts and his or her knowledge. Therefore, if he or she finds security vulnerability in some of the cryptography algorithms or transmission protocols, it would be possible to replace it.
However, there is still a question arising related to security of stored documents. And again, there are several solutions providing certain reliability level, however, in this case, TTP proves to be the simplest and the most cost effective approach that guarantees confidentiality. But we should also remind that this guaranty relates primarily to legislative, which involves of course necessity of legal support applicable worldwide, if possible.

12 Availability

There should resolutely be a distributed, self-corrective architecture. We can observe obvious inclinations to divide large data volumes into several parts and its distribution into different sites in schemes of some ongoing projects (see above). Question is, if it is appropriate to allow users to define number of instances or localities the data should occur in the whole architecture.
Another important issue we must subsequently solve is to ensure the system is aware of all existing data (and also already nonexisting data, such as in case of partial archive drop-out) and to provide some logic for simple and fast searching.
If we provide answers for the above requirements, such architecture could enable us to replace archive components in run-time, i.e. to substitute old nonsatisfying hardware with new hardware (ideally of any platform) without a need for manual data migration.

13 Non-repudiation

Even irrevocability is one of those somewhat difficult tasks. Not technically - realization is relatively clear: TTP can create trustworthy log - but objectively. If we create heterogeneous and complex architectures, after decades, we may expect development of giant log files. Discussion about issues related to log files of such system would generate enough resources for another study, well, so much about logs. However, answers to these issues are at least as tough as in case of long-term data storage.
Existence of the log itself does not assure irrevocability. The long-standing experience shows that we need human interaction in order to perform analysis, i.e. necessary element of irrevocability that is based on log files. However, regular, though just offline analysis of log (with presumed volume of tens of gigabytes per day) is impossible without automation tools (see [17]). So, there is another difficult task waiting to be solved that consists of requirements definition, chosen tool back compatibility provision and creation of so called auditing record.

14 Time Authenticity

The long-term archiving requires the data to be viewed from another perspective - perspective of time. Time of document's creation, change, storage - these all are factors carrying information that is undetachable from the document. In many cases, this information is of paramount importance, such as for dated land registry records, creation time of which represents an important attribute, or similarly for medical records etc.
Question remains, what attributes should be responsibility of long-term archive and what should be left to document owners? From conservative point of view, if someone requires evidential proof of document's creation time, it should be possible to obtain it the way this user considers to be suitable. If he or she then stores the document in a long-term archive, the archive should be capable of providing the actual insertion proof. Still, data time authenticity (i.e. provision of proof that data originated in time declared) and its long-term provision (because other time authenticity is pointless) is intimately associated with long-term data storage and represents its linchpin. Because motivation drivers for long-term data storage reside primarily in need for long-term storage of documents, such as contracts, land registry records etc., and without an option of evidential verification of its creation time, all these documents lose completely its value and infor-

mation they store.
We have already discussed time authentication methods in [11]. This discussion pointed out impropriety of the present and widespread classic time-stamp solution whereas emphasized suitable prerequisites and attributes of more challenging method - electronic notary service. Since there are no other methods of data time authentication at present besides the two mentioned, we actually considered all possibilities related to their suitability from long-term view.

15 Discussion about Security Requirements

Discussion about ensuring above components of computer security will certainly be conducted at least in two basic levels:

- By individual types.
- By determination.

The former viewpoint means that not all data types require provision of all basic characteristics. For example, cinematographic archive does not have to be confidential, not even constantly available and exposure period of its content may be much longer than exposure of, for instance, commercial contracts.
Data type also indicates data's sensitivity level. We may assume that text files will be usually more sensitive than voice or image records (of course, with an exception of corporate video records etc.)
This brings us to the latter level, which deals with question related to size of user group the data is intended for, or purpose of the data etc. For example, data intended for enlightenment of open group of users must be treated differently than data that may result, based on decision of limited number of privileged users, in far-reaching consequences.
As we can see, it is, for example, pointless to maintain all data enciphered or keep it in tens of instances and sites. Objective judgment of demands related to individual data requires knowledge of needs and opinions of its future users. This should be one of the major lines of our subsequent efforts.

16 Security Methods

All above mentioned principles are more or less determined by basic cryptography paradigms. Even if we affirm that most of usable cryptographic algorithms (symmetric or asymmetric) can be at best breached only with brute force, it doesn't nearly eliminate possibility that calculations unfeasible today in time shorter than infinity will maintain these characteristics in the future. In brief, at times of biggest fame of DES algorithm (with 56 key length), it was presumed that breach of cipher would be a long distance race. In comparison to several hours necessary for breach today, this statement may sound funny.
Breach of hash functions is also pretty frequent (SHA-0). Time necessary for obtaining sufficient capabilities in order to breach for instance algorithms based on elliptical curves in several hours is naturally still a question; however, it is obvious that issues related to long-term data storage won't accept standard used approaches.
Readers have for sure noticed that authors introduced new term - time authentication. Although this study contains its description, we recommend ([10] and [11] resources) for better familiarity with these issues.

Part VIII
The future

It's really hard to predict the next progress. But the true is that data size increases, while the place for them is limited. The logical way of solution and the well known trend is of course miniaturization (solution based on technologies). The other concept is looking for non conventional methods - e. g. data rotation in the network (solution based on innovative logic).

Part IX
Miniaturization

The miniturization is ofcourse one of the most researched category of IT. As atoms as the smallest physical components have been recovered by quarks and they have been

recovered by strings, 5.25 inches diskettes have been recovered by DVDs and they can be replace by for example "Milipede." It is a technology presented by IBM on the CeBIT trade fare in 2005. This technology is a micro electro mechanical system (MEMS). It is quite similar to a punch card, but the boxboard is replaced by little fields of polymer film, which characteristic is changed by electric current. Every field represents one bit and its radius is 10 nanometers. Fields are grouped to bigger areas, where each contains 4000 fields. The effect is very high density of records - about 1 TB per square inch. The advantages lie in the space savings and the access speed as well. All fields can be read in the same time. The problem is high error rate, which disqualifies this technology from possibilities of long term data storage.

Part X

Non conventional methods

As we mentioned above - data grids become more and more popular. The advantage could be the PVM (Parallel Virtual Machine) computing, which can be effectively implemented using grid.
A design of so called network disks could be interesting as well. This method allows the data rotation in the network, it means there is not a real hard medium for data storage. The situation is very similar to cars and parking places. There are much more cars then parking places on the world. The cause, why can all cars be used is that not all cars are parked at the same time - they rotate and just a little part of them stays. As can be seen, the twist is in using roads - communication lines. But this solution can be effective just using long bone networks with high traffic capacity, and is also dependent on their baud rates.
And many other technologies and principles are yearly presented on various conferences, but non of them still brought the real solution yet...

References

[1] Wallace C., Chokhani S., Trusted Archive Protocol (TAP), *Internet Draft, February 2003*

[2] Dockal, J., Poprocky V., Hajicekk D. C., Security Management Tools, Data Security Management October 2004

[3] Bearman, D., Documenting Documentation, Archivaria 34(Summer), 1992

[4] Codd, E. F., Relational Database: A Practical Foundation for Productivity, CACM 25(2): 109-17, 1982

[5] Michelson A., Rothenberg J., Scholarly Communication and Information Technology: Exploring the Impact of Changes in the Research Process on Archives, American Archivist 55(2), 1992

[6] Van Bogart, John W. C., Long-Term Preservation of Digital Materials, Paper presented at the National Preservation Office Conference on Preservation and Digitisation: Principles, Practice and Policies. University of York, England, September 3-5, 1996

[7] Santesson S., Polk W., Barzin P., Nystrom M., Internet X.509 Public Key Infrastructure Qualified Certificates Profile, RFC3039, January 2001

[8] Adams C., Pinkas D., Ross J., Pope N., Electronic Signature Formats for long term electronic signatures, RFC3126, September 2001

[9] Adams C., Cain P., Pinkas D., Zuccherato R., Internet X.509 Public Key Infrastructure Time/Stamp Protocol (TSP), RFC3161, August 2001

[10] Hajicekk D. C., Electronic notary services, Article Security and Protection of Information 2003, Brno, 2003

[11] Hajicek D. C., Studensky I., Time-stamp services in Czech legislation, Article (Mikulasska kryptobesidka), Praha, 2003

[12] Unzeitig J., Sykora M., Hajicek D. C., Dilci zprava z vyzkumneho ukolu Uzivatelsky pratelske videokonference, Technical Report, Brno 2004

[13] http://sitola.fi.muni.cz

[14] http://www.cesnet.cz/doc/techzpravy/2003/ibpdidas/

[15] http://www.cesnet.cz/doc/2003/zprava/medimed.html

[16] http://www.pvt.cz

[17] Dockal J., Hajicek D. C., Dilci zprava z vyzkumneho ukolu, Technical Report, Brno 2004