Integrated Information Systems – A Challenge for Long-Term Digital Preservation

Viveca Asproth
Mid Sweden University, Östersund, Sweden

viveca.asproth@miun.se

Abstract

Information technology has led to the fact that much of the information that earlier was produced on paper is now being produced in electronic form. E-government services, where the whole idea is to exchange information electronically, will radically redirect paperbound information towards electronically generated and managed information and documents. Earlier on, long-term preservation was all about paper-bound documents containing content, context and format as well as the legal issues concerning the document, such as signatures and stamps. Although paper-bound documents still are of frequent use, the degree of digital information rapidly increases. Digital information is often compiled from different databases, where content, context, format and signatures can be separated and put together differently. The challenges associated with digital preservation are strategic, organizational and structural as well as technological. The aim of this paper is to present some challenges associated with digital preservation.

Keywords: long-term preservation, digital information, integrated systems, e-services, hidden Web, inter-organizations, anticipation

Introduction

Information technology in combination with changes in organizational structures and forms for work has led to an increasing amount of information. In addition totally new forms of information and records are created. Although paper-bound documents still are of frequent use, the degree of digital information rapidly increases. Much of the information that earlier was produced on paper is now being produced in electronic form, for example as e-mail and in databases. This means that a considerably larger amount of information, that is very disparate in nature, needs to be handled. Electronically generated and stored information is also exposed to change and manipulation in quite another way than paper-bound information.

A 10-year period is a long time when it comes to preserving electronic records. This is of course due to the rapid technological development, for example the problems with obsolete software and file formats. Goldstein (2004) stresses the need for a quick solution to the problem of archiving electronic information for the future. Further, he says that the major threat to archived material today is not the fire hazard, but the rapid development of different file formats for documents, sound and images. "The last decade has generated more recorded information than any previous decade of human activity. The fact that the majority of these data is less accessible than ever before is one of the ironies of the modern infor-

Editor: Kathy Lynch

mation age" (Duranti, 2000). This implies a risk for a situation where the societal memory no longer can be guaranteed. In data mining and data warehousing the issue of archiving data is addressed, but the time horizon for this is limited and the use of archived data is for a specific purpose.

Vital functions in society such as the social insurance and national registration have to be accessible for long periods of time. Without authentic, accessible, searchable, and reliable records these functions in society are in fact impossible to maintain. The technical development and transition to electronic media has led to that traditional methods have become obsolete or at least insufficient.

Problem Domains

There still remain a lot of problems to solve before secure and efficient long-term preservation will be brought about. It seems that although much of the challenges associated with digital preservation is strategic, organizational, and structural and not only technical, the published research concentrates at solving the technical issues. Digital preservation requires, in addition to the technological development, elaborated strategies, new workflows and organizational structures, standards and common metadata, new specific competences, and close co-operation across different professions from traditional preservation management to computing science. (Asproth, 2005)

- Technical problems

Issues at stake in this domain cover hardware, software and formats. The large amount of electronic records to store, for active use, demands new mass storage media. The fragility of the media leads to low durability and there is also often a lack of knowledge about the durability of different media for storage. The software problem concerns the rapid development of new versions of the software product and occasionally changeovers to new software products. Connected to the problem with rapidly changing software, there is the problem with changing formats, which in the long run can lead to the consequence that archival information becomes unreadable (Gibbs & Heazlewood, 1999).

- Legal issues

One important requirement on records is evidence. The problem of how to ensure authenticity and how to preserve evidence is emphasized by, among others, Bearman (1994), Duranti (2001), Rothenberg (2000), and Wilson (2000). The rapidity and the ease with which changes can be made mean that there are challenges associated with ensuring the continued integrity, authenticity, and provenance of digital materials (Beagrie, 2002).

- Organizational problems

Most of the challenges associated with digital preservation are organizational – not technical. The first line defense against loss of valuable digital information rests with the creators, providers, and owners of digital information (Marcum, 1997). There is also a matter of inter-organizational problems to cope with. For example, governmental organizations want to offer electronic services to the citizens that requires collaboration as the services assume information from more than one organization (Asproth, 2006a).

- Context and metadata

How should it be possible for someone in the future to understand the context of the information and the processes that created it? This is a question of metadata, a question that is considered as more important in the digital age. How can we determine what amount of evidential historicity is required and how to capture and retain it (Bearman, 1994)?

Ongoing Research

There is a lot of ongoing research aiming to solve the technical problems associated with long-term preservation of digital information, although the projects very seldom are directed especially towards this area. Examples of research areas are mass storage media, format conversation, and migration of technical environment (hardware and software), emulation of technical environment, digital signatures and other security issues. There is also research going on to develop new models for long term preservation. As an example the Record Continuum Model (Upward, 2005) can be mentioned, which the Australian School on recordkeeping is based upon. There are two major perspectives on recordkeeping; record life-cycle model, and the Records Continuum Model (Hänström, 2005).

The life-cycle model sees Record as existing with a linear and decreasing grade of usage/activity that ends up with a decision whether the record should be destroyed or preserved. The Records Continuum Model states that records exist in a continuum space-time, and that records never are finished in their creation, they are in continuous change. The use of records could, within the continuum Model, vary and shift from inactive to active at any moment. In opposite to the life cycle model, which has its basis in paper based recordkeeping, the Record Continuum Model has its basis in both paper based and electronic recordkeeping. (Borglund & Öberg, 2006)

Current State in Swedish Governmental Organizations

To investigate the current state concerning digital information in Swedish governmental organizations a study was carried through. The study aimed to investigate how different governmental departments managed electronic information both in a short and long run. The ambition of many governmental departments is to become 24/7 agencies. Of the studied organizations, several of them have e-services taking care of at least some part of their activities. Some others plan to start e-services in a near future. All the studied organizations handle electronic records in the studied cases. They saved the electronic records in XML or other formats that is said to be appropriate for long-term preservation. Most of them also scanned paper documents and used them as electronic information. Much information is also stored as database records (Asproth, 2006b). Figure 1 shows a typical but simplified example of the diversity of information that is handled.

The active time for the electronic information varies quite a lot, but mostly the information is active for decades, in some cases for several decades. In some of the organizations not all information is preserved during that time, some sorting out is performed. Information like notes, e-mails, etc. is sorted out. As the amount of information is huge, just a selection is obliged to be preserved. However several of the organizations chose to save more than they were obliged to. Many of the organizations chose to save all the database records and did not sort out any of them. In some cases scanned documents were chosen to be the original and thus preserved. Others chose to print out the information and preserve paper documents. It deserves to be mentioned that electronic records has not existed for a very long time and a feeling of uncertainty about how to manage was expressed (Asproth, 2006b).

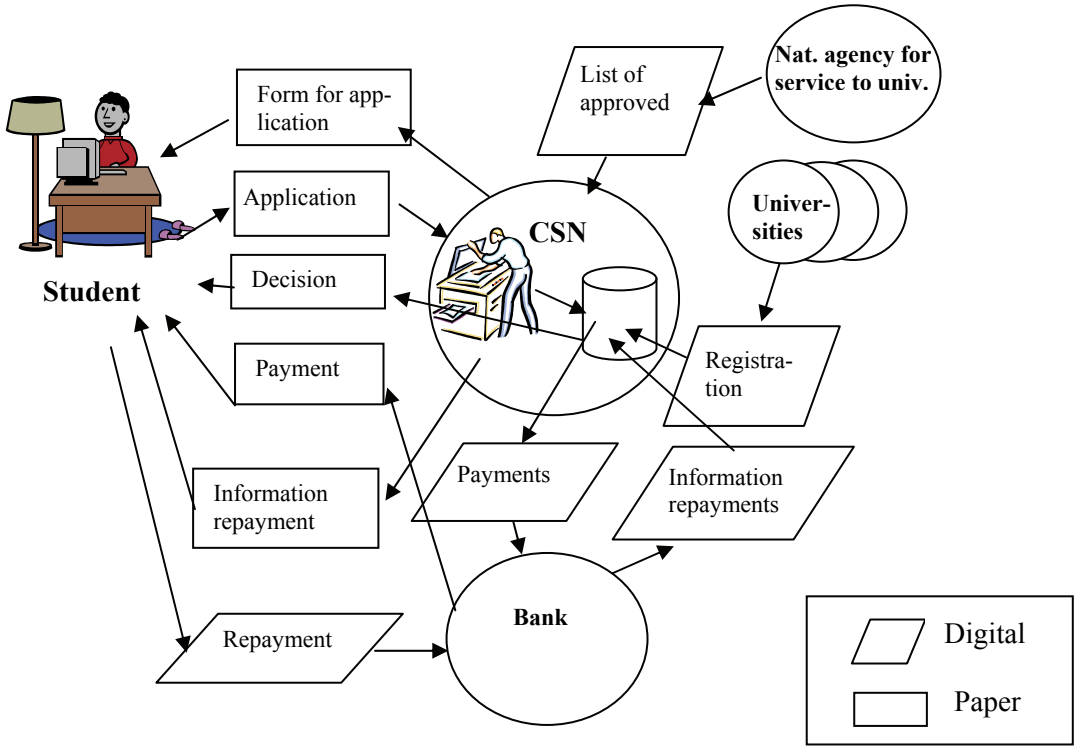

Figure 1: Information concerning students´ loans.

Integrated System and e-Services

A common problem facing many organizations today is that of multiple, disparate information sources and repositories, including databases, object stores, knowledge bases, file systems, digital libraries, information retrieval systems, and electronic mail systems. Decision makers often need information from multiple sources, but are unable to get and fuse the required information in a timely fashion due to the difficulties of accessing the different systems, and due to the fact that the information obtained can be inconsistent and contradictory. (Garcia-Molina et al, 1995)

Navathe & Donahoo (1996) expressed that the current methodologies for information integration are inadequate for solving the problem of integration of large scale, distributed information sources (e.g. databases, free-form text, simulation etc). The existing approaches are either too restrictive or do not provide the necessary functionality.

Many organizations, not least the in the Swedish governmental organizations, have large, rather old systems, which are not easy to replace due to the high investment costs. These systems were developed for their own specific purposes as the need for them occurred. They are therefore relatively heterogeneous. Some information is also managed in more than one system. Rather than to replace them with one single new system, as the development of a systems of this size would be very huge, the choice is to keep the old systems and integrate them and build new modern interfaces on top. The problem to secure the information for long-term preservation is enormous, as the information exists in such a heterogeneous environment.

The emerging Web service technology has enabled the development of Internet-based applications that integrate distributed and heterogeneous systems and processes which are owned by different organizations. Compared to centralized systems and client-server environments, the Web service environment is much more dynamic. (Yang, Lam, & Su, 2002)

We can contrast the Web with traditional, closed, deterministic multi-user systems, such as enterprise networks. Web applications involve citizen–government, customer–business, business–business, and business–government interactions. (Rezgui, Bouguetta, & Eltiweissy, 2003)

Also, service providers come and go. The users of services must have some assurances about the services and the organizations that provide the services. (Yang et al., 2002)

A number of studies (Lawrence & Giles, 1998, 1999) have noted that a tremendous amount of content on the Web is dynamic.

However, since current-day crawlers only crawl the publicly index able Web, much of this dynamic content remains inaccessible for searching, indexing, and analysis. The hidden Web is particularly important, as organizations with large amounts of high-quality information are placing their content online, by building Web query front-ends to their databases (Raghavan & Garcia-Molina, 2001).

Governments, both local and national, want to offer electronic services to the citizens. As an example can be mentioned a Swedish project to offer an e-service to the citizens for building permits. This application includes information and partial permits from several departments within the local government but also from other governmental organizations. To facilitate for the citizen the e-service will include the whole chain as one single e-service disregarding the organizational boundaries. To develop such an e-service you have to deal with the problem of integrating several separate and heterogeneous systems which are in fact dynamic and hidden. The kind of information that is dealt with here has to be preserved authentically and be reliable over a long time. There will be a need for new methods and tools for long-term preservation of information kept in disparate systems and databases, even more in an inter-organizational perspective. In research much effort is put in to facilitate the search and access to the "hidden" Web, but this will not be enough for the needs for long-term preservation.

Inter-Organizational Systems and Management

Electronic business applications are often structured by workflow declarations that span potentially numerous generic activities in different organizations (Biskup & Parthe, 2006).

E-services to citizens are fairly new and inter-organizational e-services are very unusual. Earlier all information has been paperbound and the organizations have had their "original" of the document/documents. In introducing the e-service all information is electronic and resides in the system. The problem that occurs to be solved is the responsibility for the preservation of the information, which is an organizational and even inter-organizational issue.

Development of an e-service is carried through as a project. As an e-service may live for some time and certainly the information created in the system, the case-by-case approach has its limits. (Asproth, 2006a)

Will the management forms of today's organizations make it possible to fulfill the demands of inter-organizational cooperation? Powell (1990) and Obata & Shizuka (2003) present the differences between hierarchical and network management (see Table 1).

Table 1: Comparison of Hierarchical and Network type management
(Powell, 1990) Modified 2003 (Obata & Shizuka, 2003)

	HIERARCHICAL TYPE	NETWORK TYPE
Organizational style	Layered, heavy	Flat, flexible
Normative basis	Employment relationship	Complementary strengths
Means of communication	Routines	Relational
Methods of resolution	Administrative	Norm of reciprocity
Degree of flexibility	Low	Middle
Amount of commitment Among the parties	Medium to high	Medium to high
Tone	Formal, bureaucratic	Open ended, mutual benefits
Actor preferences	Dependent	Interdependent
Resource utilization	Limited	Flexible
Job process	Reliable and rigid	Complexity
Decision	Slow	Agile

Obata & Shizuka (2003) propose to combine Network Management with Hierarchical Management for dynamism, flexible and sustainable reliance in E-business. This combining of management style called Spiral management (see Figure 2) combines relational, interdependence, reciprocity by network management and responsibility, reliance by hierarchical management.

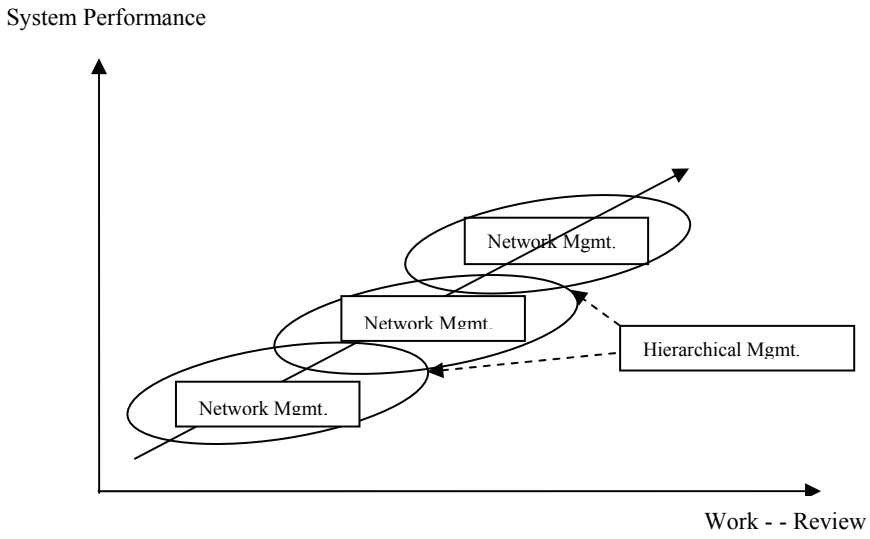

Figure 2: Spiral Management (Obata & Shizuka, 2003)

According to Zakaria, Amelinckx, and Wilemon (2004) the human challenges of virtual team membership are:

- Creating effective team leadership
- Managing conflict and global virtual teams dynamics
- Developing trust and relationships
- Understanding cross-cultural differences
- Developing intercultural communication competence

Holmqvist (2003) and Rashman & Hartley (2002), recommend organizational learning as a tool to in the first place develop an intercultural communication competence, but also as a complement learn more about each other. To develop organizational learning within an organization has shown to be successful. The question is how to transfer the concept to inter-organizations. There might be competitiveness and conflicting interest that put hindrance in the way. Another problem with inter-organizational learning is that there is mostly a case-by-case management approach. There is a need to maintain the knowledge over a longer time. People come and go in an organization and the ones that knew all about the system disappear.

An Anticipative Approach

Long-term preservation of digital information is uncertain as we can not fully foresee who the users are and for what purpose they want to use the information. Information systems up to now have been developed for a specific and clearly defined purpose. With the development of e-services the users become unknown to the systems developers and the service providers. "For example, an organization (e.g., a service provider or a service broker) cannot predetermine the users of its resources" (Yang et al, 2002). For long-term preservation of digital information yet another aspect is added, the unknown use (see Table 2). "The unknown use within the concept of records use is important because when unknown use exists, unknown users exist as well, which implicitly gives unknown requirements which affect quality that is dependent on user needs" Borglund, 2006).

Table 2: Use and users of information systems

	KNOWN USER	UNKNOWN USER
KNOWN USE	Traditional information systems	E-services
UNKNOWN USE		Long-term preserved information

It is not possible to fully know who the users are or how they want to use the long-term preserved information in the future. However, we have to act in beforehand. We are not able to *predict* but we have to *anticipate* the use. An anticipatory approach must build upon a large degree of freedom of the information. The information must be kept independent, not organized according to today's organizations. Still the provenance principle must be applied. Context information must be caught and related to the information extensively. To understand the information in the future new context information must be able to be added in the future.

Conclusions

An increasing amount of electronic information is to be long-term preserved. Many organizations are in some stage of changeover from paperbound information to electronic records. To be able to

preserve the information for a long time and to guarantee access to it in the future, the technical issues have to be solved or managed. In this field there are research on migration of technical environment (hardware and software) and emulation of technical environment – systems which run in a new operating environment but emulate a previous obsolete environment. There are also groups working with standards and formats. However, to-days systems are not built to support long-term preservation of information, nor do the existing system development methods address the issue.

Besides the technical problems, managerial issues connected to long-term preservation of digital information must be elaborated. There is very little support in the systems for management of distributed information, information that is not only distributed in where it is stored, but also in what type of information it is. New tools, able to assemble the information in a transparent way, have to be developed.

The information must be kept independent, not according to today's organization as the organizations change or even disappear over time. A new model for organization independency must be developed. These new models have to focus of the processes that create the information and the systems must be able to handle the different versions of information that these processes produce.

To maintain and develop the knowledge and to prevent conflicts and misunderstanding organizational learning is an excellent approach. New or adapted models suited for inter-organizations needs to be produced.

Furthermore the information must be kept understandable over time, which requires lot effort on context and meta-data. Supposedly new context information must be added over time. Methods for analysis of needs for new context information must be elaborated.

There is also a question about what to preserve as it is not possible to preserve everything. The legal and governance demands determine a minimum of what to preserve, but beyond that there are requirements to preserve the cultural heritage. What is a representative selection? And finally, the information must be able to be accessed now and in the future. How can the needs be foreseen or how can a completely independent access be accomplished.

References

Asproth, V. (2005). Information technology challenges for the long-term preservation of electronic information. *International Journal for Public Information Systems*, *1*, 27-37.

Asproth, V. (2006a). Inter-organizational management and decision-making. *Systemist*, *28*(2), 4-12.

Asproth, V. (2006b). Long-term preservation of electronic information – A study of seven Swedish governmental organizations. *Proceedings of the 15th International Conference on Information Systems Development*, 31 August - 2 September 2006, Budapest, Hungary.

Beagrie, N. (2002). *A continuing access and digital preservation strategy for the Joint Information Systems Committee* (JISC) 2002-2005. Retrieved November 22, 2005, from http://www.jisc.ac.uk/index.cfm?name=strategy_preservation

Bearman, D. (1994). *Electronic evidence: Strategies for managing records in contemporary organizations*. Pittsburg: Archives and Museum Informatics.

Biskup, J. & Parthe, J. (2006). Optimistic anonymous participation in inter-organizational workflow instances. In *Proceedings of the 2nd International Conference on Information Systems Security (ICISS 2006)*. Retrieved November 20, 2006, from: http://ls6-www.informatik.uni-dortmund.de/issi/archive/literature/2006/Biskup_Parthe:2006a.pdf

Borglund, E. (2006). *A predictive model for attaining quality in recordkeeping.* Sundsvall: Mid Sweden University.

Borglund, E. & Öberg, L-M. (2006). Operational use of records, In *Proceedings of IRIS 29*, 12-15 Aug, Helsingoer. Retrieved November 20, 2006, from www.itu.dk/iris29/IRIS29/6-3.pdf

Duranti, L: (2000). The InterPARES Project. *Proceedings from the February 2000 InterPARES Symposium*. Retrieved June 19, 2006, from http://www.interpares.org/display_file.cfm?doc=ip1_symposium_2000.pdf

Duranti, L. (2001). The impact of technological change on archival theory. *Archival Science, 1*, 39-55.

Garcia-Molina, H., Hammer, J., Ireland, K., Papakonstantinou, Y., Ullman, J. & Widom, J. (1995). Integrating and accessing heterogeneous information sources in TSIMMIS. *Proceedings of the AAAI Symposium on Information Gathering* (pp. 61-64), Stanford, CA, AAAI Press/MIT Press, Menlo Park, CA, USA.

Gibbs, R. & Heazlewood, J. (1999). Electronic records: Problem solved? A report on the public office Victoria's electronic records strategy. *The Journal of the Australian Society of Archivists, 27*(1), 1-13.

Goldstein, H. (2004). The infinite archive. *IEEE Spectrum*, *41*, 74.

Hänström, K. (2005). *Arkivparadigmets förändring* [The records paradigm in change] Unpublished Master thesis. Härnösand: Mid Sweden University.

Holmqvist, M. (2003). A dynamic model of intra- and inter-organizational learning. *Organizational Studies*, *24*(1), 95-123.

Lawrence, S. & Giles, C. L. (1998). Searching the World Wide Web. *Science*, *280*(5360), 98.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the web. *Nature*, *400*, 107-109.

Marcum, D. B. (1997). A moral and legal obligation: Preservation in the digital age. *International Information and Librarian Review, 29*, 357-365.

Navathe, S. B. & Donahoo, M. J. (1996). Towards intelligent integration of heterogeneous information sources. *Database reengineering and interoperability.* Plenum Press, New York, NY.

Obata, K. & Shizuka, H. (2003). Spiral management for e-business. *Journal of the Asian Design International Conference, 1* (G-01), 1-7, Retrieved October 19, 2006, from www.idemployee.id.tue.nl/g.w.m.rauterberg/conferences/CD_doNotOpen/ADC/final_paper/499.pdf

Powell, W. W. (1990). Neither market nor hierarchy: Network forms of organization. In B. Staw (Ed.), *Research in Organization Behavior*, *12*, 300.

Raghavan S. & Garcia-Molina H. (2001). Crawling the hidden web. *In Proceedings of the 27th International Conference on Very Large Data Bases (VLDB)*, pp 129-138.

Rashman, L. & Hartley, J. (2002). Leading and learning? Knowledge transfer in the Beacon Council Scheme. *Public Administration*, *80*(3), 523-542.

Rezgui, A., Bouguetta, A. & Eltiweissy, M. Y. (2003). Privacy on the Web: Facts, challenges, and solutions. *IEEE Security and Privacy*, November/December, 40-48.

Rothenberg, J. (2000). Preserving authentic digital information. *Authenticity in a Digital Environment,* 51-68.

Upward, F. (2005). The records continuum. In S. McKemmish, M. Piggot, R. Barbara & F. Upward (Eds.), *Archives: Recordkeeping in society* (pp 197-222). Wagga Wagga, Australia: Charles Stuart University, Centre for Information Studies

Wilson, I. (2000). The test of time. *Authentic records in the electronic age,* 24-34

Yang, S., Lam, H. & Su, S. Y. W (2002). Trust-based security model and enforcement mechanism for web service technology. In *Technologies for E-Services. Third International Workshop, TES 2002*, pages 151–160.

Zakaria, N., Amelinckx, A., & Wilemon, D. (2004). Working together apart? Building a knowledge-sharing culture for global virtual teams. *Creativity and Innovation Management. 13*, 15-29.

Biography

Viveca Asproth received the doctor (PhD) and the reader (Docent) degrees from the University of Stockholm. Currently she is an Associate professor in Computer and Systems Science at Mid Sweden University in Östersund, Sweden. She has published papers on visualization, spatial systems, decision support, anticipation and fuzzy systems. She is member of the board of Graduate Education and deputy head of the department of Information Technology and Media at Mid Sweden University. In her current research she is focusing on long-term preservation of digital information.