UNESCO discussion paper on digital preservation

At the General Conference of UNESCO in October 2001 a resolution was passed that draws attention to the need for preservation of the digital heritage. One of the recommendations is to draft a charter on digital preservation for the next General Conference in 2003. In preparation of this charter, the ECPA was asked to write a discussion paper on digital preservation for the Executive Board of UNESCO. This discussion paper was subsequently adapted and incorporated in documents presented to the Executive Board in May 2002.

The original version of the ECPA discussion paper as submitted to UNESCO, including useful notes and references is made available here, with kind permission of UNESCO. It is a concise text outlining the basic issues, written in non-technical language, which may serve as a useful introductory overview for management and decision makers

Preservation of digital heritage

Draft discussion paper prepared for UNESCO

Yola de Lusenet

European Commission on Preservation and Access

ecpa@bureau.knaw.nl

March 2002

1. Introduction and executive summary

The vast amounts of information produced in the world are now for a large part digital and include a wide variety of materials: text, databases, audio, film, images. [1] They range from medical records to movie DVDs, from satellite surveillance data to websites presenting multimedia art, from data on consumer behaviour collected by supermarket tills to a scientific database documenting the human genome, from news group archives to museum catalogues.

For the cultural sector, traditionally entrusted with collecting and preserving cultural heritage, the question has become extremely pressing as to what of this enormous amount of materials should be kept for future generations, and how to go about selecting and preserving it.For conventional materials, a certain amount of agreement exists on how to collect information worth preserving. Legislation is in place, policies and procedures have been developed, and all the parties involved -the private sector, government agencies, academic institutions, archival institutions, libraries and museums- are carrying out their respective roles.

With the advent of digital media, however, a new and complex environment has come into being. Not only the media are new, the contents and the means of distribution have also changed dramatically, and new players –among users as well as creators of information- have entered the stage. With so many new types of content presented in new ways, it has to be redefined what is worth keeping and how this can be achieved.

The speed at which things digital move has upturned the order of all established preservation practices. Generations of platforms, programmes and machines succeed one another so quickly that it is a matter of years rather than decades before materials become inaccessible as a result of compatibility problems. The timescale for preservation has shrunk: steps to ensure that digital materials remain accessible have to be taken very early on in their lifecycle. At the same time, to keep them alive over time, preservation has to be a continuing process of periodical activities to keep up with an environment constantly in flux.

Preservation of digital heritage requires proactive strategies and a cooperative effort by both the producers and keepers of information. The support of (national) policies and legislation is needed to enable heritage institutions to meet these new challenges. New frameworks have to be created, tasks and responsibilities redefined, new forms of cooperation established. Technological problems are huge and only partly understood, and issues of rights and ownership more complex than ever before.

The rapid spread of information technology makes preservation of digital heritage a worldwide concern. Most of the digital information is as yet produced in the Western world, but other countries will soon catch up. More and more digital systems for administrative purposes are being introduced everywhere, and a great many countries are digitizing cultural materials for better access. The problems that some are facing today, others will have to deal with tomorrow –and in the digital world, tomorrow comes more quickly than you think.

Governments and policy makers should be aware that preservation of digital heritage is an urgent issue and that solutions cannot be found overnight. The risk of losing essential materials in which valuable resources have been invested is very real. It is therefore crucial that countries assume responsibility for digital heritage and take steps to prevent such loss.

2. Existing models and legal frameworks

Traditionally, preservation of cultural heritage has been supported by legal frameworks and procedures which are largely based on formal criteria. National libraries collect and preserve publications through legal deposit of the national production, and there is extensive archival legislation defining when and how records must be transferred to archives for selection and preservation. Specialized archives and museums have responsibilities for collecting and preserving sound, photographs or film. Legislation may vary considerably between countries (e.g. regarding the categories of material to which legal deposit applies), but there is wide agreement on the basic principles, and all parties involved in the process are well aware of them.

In the digital world, new types of materials have come into being that are hard to classify by conventional criteria. Multimedia materials combine different types of content with different functionalities. A database is not a fixed object that can be stored in a definitive form, nor can one separate the data from the relationships between them. Websites may combine files with various types of content –data, texts, images, sound– and many of them are (partly) dynamic. Websites may also be distributed sites including materials stored on different servers at different locations in the world. Such mixed or dynamic materials do not fit into traditional categories; on the basis of existing policies it is often not possible to decide where the primary responsibility for collecting and preserving them should lie.

Especially on the web, the usual filtering mechanisms of publishers and other agencies that review and select materials worth publishing often do not apply. Institutions and individuals now publish their own materials, which may be just drafts, or different versions of texts published also elsewhere. Although we speak of '*publishing* on the internet', it is not at all clear what constitutes an internet publication. Place of publication, an essential criterium in deposit legislation, can no longer be used to define the national production or imprint: domain names do not necessarily reflect where the material is produced and in which language, and many sites are mirrored in other locations.

That raises the question of which materials should be considered publications as defined by deposit legislation, and how deposit legislation can be adapted to include digital materials that national libraries should preserve. Although some countries have extended legislation to cover offline publications such as CD-ROMs, the case of online materials is as yet still diffuse. [2]

In the archival world, electronic records have taken the place of paper. In the absence of a fixed object that can be preserved as is, it becomes necessary to decide which elements actually make up an authentic electronic record. With records being used for years or even decades, they will inevitably have to be moved from outdated environments to new ones, with the risk of changes or loss of content, functionality or original appearance. As it will almost certainly be impossible to keep the original record as it was once created, it has to be determined which are the significant properties of records that must be preserved in order for them to function as authentic, meaningful records over time [3] . With many new types of material appearing on internet sites, it also has to be established to which extent these should be regarded as records covered by archival procedures and legislation. [4]

Legal frameworks defining responsibilities and procedures need to be adapted or extended to be able to deal with the new digital environment. Adequate legislation in this area is a necessary instrument for institutions to define tasks and select materials for preservation.

3. The internet as a cultural space

The internet consists of one billion pages and keeps growing. A number of these pages are devoted to materials of the kind that we traditionally associate with heritage institutions: electronic journals and articles, newspapers, photographs, catalogues and finding aids, and other information and documents from the public sector.

However, the internet is an extremely democratic medium, and on the other end of the scale there are innumerable websites created by individuals and informal groups. Virtual communities of people scattered over the globe but united by shared interests discuss just about anything under the sun, including such topics as endangered languages or regional cooking. Artists experiment with multimedia websites as new art forms, amateur genealogists present data on their family's history. Taken as a whole the internet in many ways reflects our society, as a huge open space where a variety of cultural activities are pursued.

Preservation of digital heritage will somehow have to deal with new manifestations of cultural content on the web, which challenges traditional classifications of materials worth keeping. Unfortunately, it is risky to rely on time to sift what may prove to be of lasting value from the merely ephemeral. Websites are changed and updated constantly, and superseded materials vanish without leaving a trace. Estimates for the average life expectancy of a webpage vary from 44 days to two years. [5] When organizations go out of business or lose interest, whole websites disappear from sight.

This not only happens with informal or temporary sites, but also with central and official ones - like the White House site, www.whitehouse.gov, which was wiped clean when George Bush took over the presidency. The collection of speeches and official communications of the Clinton administration disappeared overnight, breaking a massive amount of links to these materials on other sites [6] . Luckily, the Clinton materials have been partly saved by the National Archives and Records Administration (NARA), which had archived several versions of the site across the Clinton years.

NARA is among the heritage institutions that, recognizing the risks posed by the instability of the internet, have opted for a proactive approach. From the wide diversity of materials on the web, they aim to safeguard access to what is potentially of lasting cultural value. However, their work is complicated by the fact that there are no established formal criteria to select websites for preservation. As webmaterials cannot easily be classified into well-known categories, new policies need to be developed to ensure that all kinds of webcontent that may be of value for later generations are indeed saved for posterity.

4. Approaches to digital preservation

4.1 *Science data*
A number of initiatives to preserve digital materials have been ongoing for some time.In scientific and scholarly research, computerized data have been created and used for decades. The space and earth observation communities, using massive amounts of data that need to be studied over a long period of time, have been very active in developing a reference model for archiving data that is being widely adapted [7] . Data archives, especially in the social sciences and the humanities, have for years been collecting data sets created in research projects so that they are maintained and can be re-used [8] .

4.2 *Library initiatives*
National libraries generally approach the digital environment from the angle of deposit legislation. Deposit of offline digital products, such as CD-ROMs, is in several countries already a legal requirement [9] . Online electronic journals are treated as an extension of a long tradition of print publishing, which libraries have always collected and preserved. To ensure continued access to the whole of the scientific electronic journal environment, including live links, data and multimedia presentations, libraries are now trying to come to arrangements with publishers about deposit, as yet often on a voluntary basis; the first such agreement was drawn up by the Royal Library of the Netherlands with the Dutch Publishers Association [10] . As it concerns paid publications with restricted access, cooperation of publishers is essential, and among publishers awareness that provisions have to be made for long-term archiving is growing. A project funded by the Mellon Foundation aims at an active partnership between academic publishers and large research libraries to create e-journal archives [11] .

Several libraries have developed strategies for actively selecting and preserving websites on the basis of a concept of 'publication' [12] , of which the Pandora project of the National Library of Australia is perhaps the best-known example. 'Publication' is defined in broad terms: anything on the internet is regarded as a publication, only organisational records are explicitly excluded [13] . At the center of the policy is the idea of national production constituting national cultural heritage: sites selected for preservation should be about Australia, or deal with a topic highly relevant for Australia and written by an Australian. Selection is determined by content and 'high priority is given to authoritative publications with long term research value'. [14]

4.3 *Initiatives by archival institutions*
Some national archives, as for instance the Public Record Office and the National Archives of Australia, have extended policies for electronic record management to include websites of government agencies (public sites as well as intranet sites) and developed guidelines describing best practices. The Public Record Office warns that materials on websites are not always recognized as records. They are often 'very different in nature from the traditional image of a "record". So much so that it can tend to give the impression that no records are present. This can be highly misleading' [15] . For, on the contrary, 'rigorous records management' is required also for websites. Responsibilities and procedures for identifying records and managing them remain valid in the internet world.

Other institutions are focusing on collecting materials in a specific discipline. The International Institute of Social History, as a research institute with the task of collecting and archiving materials relating to social history, decided in 1994 to collect internet documents on politics, social affairs and ecological issues. Their collection policy is exceptional in that it also includes newsgroups, and they have now collected 900,000 messages from 974 newsgroups which are accessible over the internet [16] .

4.4 *Harvesting the web*
Apart from these selective approaches for preserving webcontent, there are also examples of comprehensive approaches, which collect enormous numbers of webpages without any selection for content. The Internet Archive, started in 1996 as a private, nonprofit enterprise, 'is working to prevent the Internet — a new medium with major historical significance — and other "born-digital" materials from disappearing into the past' [17] . It collects freely available

webpages worldwide and now comprises over 10 billion webpages or 100 terabytes of data (5 times the size of all the materials held by the Library of Congress). The Internet Archive launched a 'Wayback Machine' in October 2001 to provide free access to the archive over the web [18] .

In Sweden, the Kulturarw3 Heritage Project has been harvesting Swedish websites since 1996. Robots capture all websites in the Swedish domain or providing content on Sweden [19] ; access is possible via the web. In the Finnish EVA project all 'the freely available, published, static HTML-documents with their inline material like pictures, video and audio clips, applets etc' [20] in the .fi domain are harvested. This activity of harvesting all materials freely published in the Finnish internet is regarded as complementary to the legal deposit of paid materials by established publishers.

At the moment, the main aim of these initiatives is to save webmaterials that would otherwise in any case have been lost forever. They give us 'both a record of the origins and evolution of the Internet, as well as snapshots of our society as a whole around the turn of the century' [21] . However, rendition of captured sites is as yet incomplete, for capturing online information is extremely complex. Links to external sites will in many cases be broken and interactive navigation cannot always be retained. More and more webpages are dynamic, generated 'on the fly' by databases hidden behind the static front end of the site. It is estimated that the databases behind websites, together called the 'deep web', contain many more times the amount of information accessible on the surface . The information in those databases cannot be captured by copying the website, as it is not available in ready-made pages at the surface. [22] Moreover, capturing webcontent is only the first step in a preservation process. After only five years of archiving, there is no saying yet how it can be ensured that these materials will still be available after 25 or 50 years.

In spite of many uncertainties, the initiatives taken by memory institutions are valuable explorations of the legal, organizational, economic and technical frameworks required for preservation of on- and offline materials. The experience gained by the pioneers in this area will be of huge benefit to the whole cultural sector and will contribute considerably to the development of infrastructure and policies for preservation.

5. What is preservation of digital heritage

5.1 *Artefact, information, functionality*

In the world of print, preservation can be achieved by preserving the paper object or, if that is not feasible, creating a durable surrogate for instance on microform The equivalent in the digital world would be, for example, to preserve a CD-ROM, or transfer its contents to another type of carrier. However, this does not achieve much more than preservation of the actual bits that make up a file. Whereas this is obviously a necessary condition for preservation, it is not sufficient to ensure that the information can be read and interpreted in the long run.

As file formats and programmes also become outdated, preservation of digital materials has to deal not only with maintenance of the files themselves but also with ways of keeping them accessible. This means that either the programmes have to be preserved as well and somehow kept running on new platforms, or the files have to be converted to another format that can be interpreted by new programmes. As the digital world moves on all the time, this is a continuous process if materials need to be kept accessible for decades (or even forever). In many cases this will, sooner or later, result in loss of information, functionality and/or appearance, especially with complex, multimedia materials that combine a variety of file formats and applications.

This poses risks for integrity of digital materials: how can it be ensured that the digital object, moving from one environment to the next, remains complete and undamaged? A different but related issue is authenticity, which relates to the trustworthiness of materials, in particular of electronic records. As records are used for accountability and as evidence of transactions, it is crucial for future reference that the original exists as it was first created and that the record indeed is what it purports to be. Integrity and authenticity do not only depend

on protecting files against intentional changes by unauthorized persons, but also on controlling inadvertent changes resulting from mis-interpretation or mis-representation by computer systems. [23]

5.2 *Aim of preservation*
Because carriers are temporary and environments change, preservation of digital materials cannot be understood in terms of fixed objects that should be kept in their present form. It is first of all a matter of defining the content and properties that need to be represented in future systems. In the case of electronic records, 'presentation', i.e. appearance, is not considered essential as 'it depends largely on the medium used to display it' [24] . Data in a complicated tablemay be 'frozen', i.e. only the results of the calculations are kept, not the software to produce them- or they may be kept 'alive', by retaining the software, thus offering future users possibilities for searching, selecting and sorting. [25]

If optimal functionality and access is the primary goal, it may even be necessary to upgrade to future requirements and devise systems that can incorporate the improvements of developing technology. Otherwise, future users will have to be satisfied with a level of access and functionality limited to what was possible in days (then) long gone. The present-day equivalent would be that 'access to old records required entering queries on punch cards, in FORTRAN or COBOL, with output limited to printouts in upper case'. [26]

In contrast, if there are reasons for representing materials in a historical context, one may wish to retain as much as possible of the original, so that future users can experience the material as we experience it now. These issues come up in the preservation of electronic art as for some artists the way the work is displayed (e.g. on a specific type of screen or using a specific browser) is an integral part of the work. To ascertain what the work really is and how it is meant to be shown, museums now often collect information on artists' intentions to guide preservation efforts [27] .

5.3 *Documentation of the object*
As the aim of preservation varies, so will the requirements for future representation and consequently the technology to meet them In all cases, adequate representation at a later stage depends on the identification of the type of content and file formats as well as the software that makes access possible. Only if one knows what one is dealing with can suitable preservation measures be taken. Documentation starts at the lowest level, by describing the characteristics of the bit stream as well as the hardware/software environment capable of rendering the object in its present form [28] .

Additional documentation is needed to understand and evaluate what is presented: information presented as such, without context and background information, will be hard to 'place'. Especially for electronic records, the context of materials is of crucial importance. It makes all the difference for understanding a map with red dots on it whether it was used for geological exploration or military actions, and this cannot always easily be seen from the map itself if it is presented in isolation. It therefore needs to be specified how and when the material came into being, who has held it, and how it relates to other information. Documentation also needs to include data on changes made over time, transfer from one format to another and on with authenticity (e.g. by using checksum or digital signatures). [29]

Preservation of digital materials requires first of all definition and description of intellectual content: what the material is, how it is meant to be presented and what it is meant to do. The choice of technological solutions depends on the requirements for future representation: 'any technological choice we make has inescapable implications for what will (and will not) be preserved. In the digital case, we must choose what to lose' [30] . Documentation of materials is a prerequisite for understanding how they are meant to be preserved, and constitute a considerable additional burden on heritage institutions. To facilitate preservation, efforts will have to concentrate on developing standards for documentation for specific classes of materials and on exploring how processes can be partly automated.

6. Technological issues

Most digital materials cannot meaningfully exist outside the digital environment as they rely on software for their interpretation and functionality. Printing the information out on paper to preserve it would only work for a small category of straight text files. Generally, in order to use the material at some future moment as it is meant to be used, both content and functionality need to be preserved. [31] In many cases, the 'look and feel' of the material is an aspect that cannot be ignored either. Preservation of digital materials is therefore a complex technological task that has to deal with several aspects simultaneously.

6.1 *How materials become inaccessible*
Basically, there are three ways in which digital materials can become inaccessible: (1) degradation of the media on which they are stored, (2) obsolescence of software making it impossible to read digital files, and (3) introduction of new computer systems and peripherals that cannot handle older materials.

Tapes and disks are all subject to physical decay and none of them has a lifespan that is comparable to that of preservation-standard microfilm or acid-free paper. They need to be stored under controlled conditions, but even so materials should be copied onto new media at regular intervals to prevent loss through deterioration of carriers. [32] 'Refreshment' of materials, i.e. transferring them to new media, often becomes necessary because a specific type of disk or tape can no longer be used in current computer systems. The disappearance of the 5 1/4 disk and the accompanying disk drives is a case in point. Refreshment is a recurring activity in any preservation programme.

In fact, the media on which information is stored are transitory carriers that serve their function only for a limited period of time, and preservation has to take account of other aspects as well. Obsolescence of software and hardware leads to (partial) loss of information or functionality of files in their original format. Successive versions of programmes may be compatible, but software producers do not usually support compatibility over a long period. Programmes disappear from the market or can no longer be used on a new platform. The combination of dependence on old versions of programmes that used to run on old platforms of outdated computer systems inevitably spells digital death.

6.2 *Technological approaches*
For the short term, it may be possible to keep the original environment (hardware and software) functioning. There is, however, wide agreement that this will not work in the long run, as it will result in an ever-growing collection of outdated computers and peripherals that is very hard to maintain over time. Such computer museums may still have a role for exceptional cases.

Different approaches have been suggested to combat obsolescence of software and hardware. One method is to convert files to new platforms or different programmes. This is especially attractive if they can be converted to a standard, nonproprietary format, as this facilitates maintenance over time. However, conversion may lead to unacceptable loss of functionality, especially with complex databases or multimedia materials. Even with relative simple materials it is hard to predict the cumulative effect of successive conversions over time [33] .

Other approaches aim to recreate superseded versions of operating systems and programmes in new environments, so that the files can be kept in their original format and read with the software in which they were first created. This is certainly a way to bridge one or two generations of platforms, but over time, as new systems keep being introduced, one may be faced with a Chinese boxes effect that becomes complex to handle. Another disadvantage may be that functionality is kept at the level of outdated systems, which may not be very satisfactory for future users.

6.3 *Standards and documentation*
These approaches are not mutually exclusive but should be combined in an institutional preservation policy [34] . It is as yet uncertain what will prove to be feasible and successful, and many institutions are doing research, creating testbeds and pilots to gain more experience with potential solutions. In the meantime, a better appreciation of the risks and

complexities among producers of digital materials could make all the difference for institutions engaged in developing preservation systems.

Producers can facilitate preservation efforts by using (official or *de facto*) standards. Emerging standards like XML and TIFF are promising because they are open standards not dependent on a specific platform; others, like PDF, are so widely used that this offers some hope that they will be supported over a long time. The use of proprietary software complicates matters not only because it is protected, but also because it is often inadequately documented. Even when programmes are taken off the market, source codes are not usually brought into the public domain. Adaptations made during the life of the software are not always documented, so that one cannot predict the outcome of a conversion in every detail [35] .

Creators of digital materials and the ICT industry have to be involved in the process of preservation as their cooperation can reduce the burden for heritage institutions. Creators will have to be encouraged to use open standards and provide adequate documentation of files. The ICT industry should be convinced of the value of open source software and of the need to publish detailed and complete documentation, to make sure their products can continue to be used in a preservation setting.

The technology for preserving digital materials requires investments in research and development that are substantial. However, such investments are negligible compared to the resources invested in creating the materials in the first place, and the cost to society if no adequate systems are developed and materials are thereby lost.

7. Organizational issues and responsibilities

7.1. *Responsibilities of creators*

Traditionally, the roles of creators and of keepers of information have been quite distinct. Basically, those who created materials had no interest in their preservation, and those who kept materials had no control over their creation. This division of tasks has to be abandoned in the digital world, where things move so quickly that 'digital information lasts forever –or five years, whichever comes first' [36] . Preservation requirements have to be taken into account very early on, even at the point when material is created, and 'the first line of defense against loss of valuable information rests with the creators, providers and owners of digital information' [37] .

Creators should be made aware that choices made at the time of creation affect the possibilities for later archiving. The use of standards and open formats, adequate description and documentation, and the use of permanent names for online resources facilitate long-term preservation and help to reduce costs. [38] Creators should realize how good practice in producing digital materials can help to maintain them over time.

Many producers of information manage their own materials for some considerable time after they have been created and in doing so will have to deal with preservation-related issues. Record-creating agencies often have to retain records for decades and have to make sure they can be accessed and used: in the past, national archives were expected to take preservation measures for records which they received only after twenty or thirty years [39] .

For publishers management of their own materials for business purposes has also resulted in an involvement in preservation. They are motivated to keep digital materials accessible, often storing them in standard formats such as SGML and XML, because it is commercially attractive to be able to re-use them for new products. Another consideration is that for libraries subscribing to electronic journals access to backvolumes is of crucial importance, also in case they decide to discontinue their subscription. As libraries do not physically hold the e-journals to which they subscribe, they depend on publishers for such continued access to older materials. Leading academic publishers like Elsevier have recognized that they have a responsibility to guarantee continued access and are developing their own archiving systems [40] . At the same time, the publishing industry underwrites the role of libraries and relies on them for long-term preservation. A joint draft statement of IFLA and IPA explicitly

distinguishes short-term archiving by publishers (for as long as publications are economically viable) and long-term archiving by libraries [41] .

7.2 *Cooperation*

The cooperation of creators and owners of information in developing working models for preservation is crucial. For instance, copyright issues need to be resolved before libraries can take any steps to maintain materials. Copyright legislation places such strict limitations on copying that even transferring files to the library's system may constitute an infringement of the rights of owners and creators. Although publishers recognize that copyright may be a barrier for long-term preservation, they are at the same time wary of any arrangement that would interfere with their commercial interests, by making deposited materials easily accessible on networks.

There are some examples of arrangements between libraries and publishers that aim to balance the interests of both parties, allowing copying only for preservation purposes while restricting access [42] . However, rights management is developing into an extremely complex area and not all aspects can be covered by agreements between publishers and libraries. When a digital product relies on proprietary software owned by third parties, the creator of the content does not usually hold these rights. Software vendors have so far hardly been involved in preservation efforts and software is not usually covered by deposit legislation [43] . A dazzling array of rights may be associated with websites combining mixed materials from various sources. Agreement on the principle of the right to copy for preservation will therefore have to be sought to make copyright aspects of preservation more easily manageable.

Recognition of the different interests of all parties concerned lies at the basis of models defining their roles. Ideally, responsibility for preservation is shared by creators and keepers, each maintaining materials during a certain stage of their life cycle. However, there is a risk that creators are not sufficiently aware of the need for continuing maintenance. As neglect may easily result in loss of materials, heritage institutions actively seek their cooperation and provide guideance on creation and preservation. Deposit regulations should help to ensure that materials are indeed transferred to an archiving institution. Such regulations should not only be developed for records and publications, but also for instance for research data, by making deposit a condition of research grants.

7.3 *Infrastructure for digital archives*

Already in 1996, it was proposed to build a 'deep infrastructure capable of supporting a distributed system of digital archives' [44] . Such a system would depend on trusted organizations capable of keeping materials alive for the long term and making them available to users as agreed with the depositor. National libraries and archives are at present taking on this role. For instance, the NEDLIB project [45] , led by the Royal Library of the Netherlands, explored how the Open Archival Information System (OAIS) Reference Model [46] could be used to create a networked European deposit library. There are, also a number of specialized research institutes and data archives that clearly see a role for themselves. Apart from those already actively involved, there is a range of other institutions that may have a task in preserving certain types of materials (digital photographs, sound, art, broadcasting materials) or preserving materials for a specific community (institutions with a local or regional task, research institutes in a specific discipline). A distributed system of digital archives makes it possible for institutions to specialize, by focusing on specific types of materials or on serving a specific community.

Digital archives need to be trusted organizations. Those who transfer materials for preservation have to be certain that integrity and authenticy are ensured, that technical measures are taken in time, and that rights and restrictions for access will be observed. At the moment, tasks and responsibilities of such trusted repositories have not been defined [47] . National institutions that have long been entrusted with preservation tasks are in the process of testing models for long-term preservation. Their leadership can help other heritage institutions to understand the requirements for an operational preservation system and to set up systems for their own field.

Preservation of digital heritage is as yet an unknown territory for most institutions. When they take on responsibilities in this area, they will have to adapt organizational structures and redefine tasks of staff. Cooperation and exchange of experience will be essential to avoid expensive mistakes, and training programmes for staff are a priority for all institutions facing the digital challenge.

Cooperation, guideance, leadership and sharing of tasks are all key elements of programmes for preservation of digital heritage. Cultural institutions need the cooperation of creators of information and of software producers. The creation of a system of distributed archives depends on national guideance as well as international cooperation. However, the terrain is so new and experience as yet so limited, that immense efforts will be needed to build up the necessary infrastructure. Adequate resources and support at policy level are indispensable to ensure that future generations will still have access to the wealth of digital resources in whose creation we have invested so much over the past decades.

8. Elements for a charter

1. **A large part of the world's information is now produced digitally, and most of this exists in digital form only. The web functions as a resource for information and communication as well as a cultural space where a diversity of materials are produced that cannot easily be classified in well-known categories. Much of these digital materials is potentially of lasting cultural value, and new, pro-active strategies need to be developed to ensure it is saved for posterity.**

2. **Preservation of digital heritage is an ongoing activity that requires commitment and involvement, not only from heritage institutions, but also from governments, policy makers, producers of information, and the software industry. Solutions depend on large-scale cooperation and the creation of a lasting infrastructure.**

3. **A clear division of tasks and responsibilities, based on existing roles and expertise, needs to be established. To come to an infrastructure of distributed archives, it has to be determined which requirements organizations should meet to function as trusted digital repositories. It should be established how tasks can be shared between national heritage institutions and discipline-oriented organizations working (internationally) for specific communities.**

4. **Existing legislation should be adapted to support national heritage institutions in the preservation of digital materials. Deposit legislation should extend to all materials regarded as publications, and legal frameworks for archives should include everything that constitutes a record, in whatever format it was produced. Additional procedures will have to be developed for materials that fall outside these categories (such as research data).**

5. **Awareness of preservation issues should be raised with producers of digital information. They should realize the importance of the use of standards and open source software and of adequate description and documentation. Outreach strategies of heritage institutions are needed to provide guidance and establish strong cooperation with the creators of materials.**

6. **The ICT industry should be made aware of the need to take preservation requirements into account. The value of standards and open source software should be promoted among software developers. They should be encouraged to make detailed and complete specifications of their products publicly available, especially for (versions of) programmes no longer on the market. Initiatives should be developed to build sustained repositories of specifications, documentation and related software.**

7. **Copyright legislation should not act as an impediment for preservation of digital heritage. Owners of rights, of content as well as software, should be**

convinced of the need to allow heritage institutions to take actions necessary for preservation of materials. It should be possible to carry out such actions in the framework of general agreements specifying conditions for access and use.

8. **The leadership role in digital preservation of a number of heritage institutions worldwide should be acknowledged. Their pioneering work in exploring legal, organizational, technical and economic aspects can provide the basis for defining best practices which should be strongly promoted in the whole community.**

9. **Further research to develop promising models and technology should be widely supported in order to come to fully operational systems for preservation of digital heritage as quickly as possible. As the digital world moves ahead at a rapid pace, the risk that materials will be left behind and irretrievably lost is very serious. With so many resources being invested in the creation of digital materials, it is crucial to stimulate efforts aimed at keeping them accessible also in the future.**

10. **Extensive training programmes are needed to disseminate the expertise and experience gathered so far widely among management and staff of heritage institutions. Preservation of digital heritage requires new organizational structures, new approaches and new ways of thinking. Programmes will have to focus, not only on technical aspects, but also on training staff to deal with a changing environment and new directions.**

Notes

[1] It has been estimated that the annual production of unique information is the equivalent of 1-2 billion Gigabytes, of which more than 90% is actually stored in digital form, and only 0.003% is originally produced in print format. However, in evaluating these figures one should remember that quantative measurements in terms of storage space do not say anything about the relative 'weight' of contents: one photograph easily take as much space to store as several volumes of text in ASCII format. 'How much information? http://www.sims.berkeley.edu/how-much-info

[2] See http://www.nla.gov.au/padi/topics/67.html and http://www.kb.nl/coop/nedlib/results/local_situations_v2.htm for comparisons of deposit legislation in different countries. Examples of countries where steps have been taken to include online materials in deposit agreements are for instance Denmark (see Henrik Dupont, 'Legal deposit in Denmark –the new law and electronic products', *LIBER Quarterly, the Journal of European ResearchLibraries*, Vol. 9:2, 1999. http://www.kb.nl/infolev/liber/articles/dupont11.htm) and The Netherlands (http://www.kb.nl/kb/dnp/overeenkomst-nuv-kb-en.pdf).

[3] 'With the rapid development of information and communications technology in government a wider range of record types will emerge: website (hypertext) documents, multimedia documents, digital audio and video, and dynamically interlinked documents. Many of the developments in desktop information technology will tend to blur the boundaries between types of records, and increase the problems of capturing and retaining all elements

of a record.' http://www.pro.gov.uk/recordsmanagement/eros/principles0.htm.. For an extensive discussion of principles, see InterPARES *Preservation Task Force Final Report*, Draft October 2001, http://www.interpares.org/documents/ptf_draft_final_report.pdf and Jeff Rothenberg and Tora Bikson *Carrying Authentic, Understandable and Usable Digital Records Through Time*. Report to the Dutch National Archives and Ministry of the Interior, 1999.

[4] See e.g. *Archiving Web Resources: A Policy for Keeping Records of Web-based Activity in the Commonwealth Government*, National Archives of Australia, revised version January 2001, http://www.naa.gov.au/recordkeeping/er/web_records/policy_contents.html**,** and *Management of Electronic Records on Websites and Intranets:An ERM Toolkit*, Public Record Office, December 2001, http://www.pro.gov.uk/recordsmanagement/eros/website_toolkit.pdf

[5] Anne R. Kenney et al, 'Preservation risk management for web resources'. *D-Lib Magazine*, Volume 8:1, January 2002. http://webdoc.sub.gwdg.de/edoc/aw/d-lib/dlib/january02/kenney/01kenney.html

[6] Richard Wiggins 'Digital preservation: paradox and promise'. *Library Journal/netConnect* Spring 2001, http://www.libraryjournal.com/digital-preservation.asp

[7] The Consultative Committee for Space Data Systems is coordinating the development of the Open Archival Information System (OAIS) Reference Model which is now a draft ISO standard. (http://www.ccsds.org/documents/pdf/CCSDS-650.0-R-2.pdf)

[8] Social science data archives in Europe are listed at http://www.nsd.uib.no/cessda/europe.html; in the UK for instance there are also several specialized repositories, like the Archaeology Data Service collecting digital records from excavations (http://www.ads.ahds.uk)

[9] For instance in France, Norway, Canada, Germany and Austria; for discussion and overview see http://www.nla.gov.au/padi/topics/67.html

[10] See http://www.kb.nl/kb/dnp/overeenkomst-nuv-kb-en.pdf; other examples: Australia (http://www.nla.gov.au/policy/vdelec.html) and the UK (http://www.nls.uk/professional/legaldeposit/nonprint/code.html)

[11] It concerns a group of six research libraries and six major academic publishers, cooperating on several projects. See Dale Flecker, 'Preserving scholarly e-journals', *D-Lib Magazine*, Volume 7:9, September 2001. http://www.dlib.org/dlib/september01/flecker/09flecker.html

[12] E.g. the Bibliothèque nationale de Québec has formulated the principle that 'networked publications are as important as traditional publications' and initiated a programme that is consistent with their legal deposit programme which assumes that networked publoications, just like traditional publications, will be deposited at the beginning of their 'active life'. Selection includes 'independent, coherent publications, mongraphs, serials' and excludes for

instance institutional websites 'taken as a whole'.
http://www.bnf.fr/pages/infopro/ecdl/quebec/sld027.htm

[13] 'The NLA is operating on the basis that anything that is publicly available on the Internet is published. However, distinctions between traditional categories of documents such as books, serials, manuscripts, working drafts and organisational records are blurred in the electronic environment. It is not the intention of the Library to preserve organisational records and similar materials, which are the domain of archives and record management.' *Selection Guidelines,* 3.3, http://pandora.nla.gov.au/selectionguidelines.html

[14] .http://pandora.nla.gov.au/selectionguidelines.html

[15] *Management of Electronic Records on Websites and Intranets:An ERM Toolkit*, Public Record Office, December 2001, p.7,
http://www.pro.gov.uk/recordsmanagement/eros/website_toolkit.pdf

[16] http://www.iisg.nl/occasio/index.html Even though newsgroup messages are straightforward ASCII texts, the process of storing them has turned out to be 'surprisingly complicated', primarily because of fragility of storage media (http://www.iisg.nl/occasio/Occasio-uk.PDF)

[17] http://www.archive.org/about/index.html

[18] http://www.archive.org/index.html

[19] Different categories are collected: (1) addresses ending in **.se** (2) web servers located in Sweden but whose addresses end in **.com, .org,** and **.net** (3) pages by Swedish producers on a foreign server (very popular: the **.nu** domain of Niue, a small nation in the Pacific; *nu* is Swedish for 'now'), and (4) *suecana* –pages abroad with content about Sweden, e.g. travel or translations of Swedish literature. http://kulturarw3.kb.se/html/projectdescription.html

[20] Databases and programmes are excluded. Kirsti Lounamaa and Inkeri Salonharju, 'EVA - The acquisition and archiving of electronic network publications in Finland', *Tietolinja News* 1, 1999. http://www.lib.helsinki.fi/tietolinja/0199/evaart.html

[21] Brewster Kahle, founder of the Internet Archive,
http://www.archive.org/wayback/press_kit/press_release.html

[22] One could theoretically copy both the database and all the software necessary to generate the pages and in this way preserve the functionality; however, as the programmes run on the server, one would need to gain direct access (as reserved for the webmaster) in order to copy them.

[23] 'The overall process of preservation must be continuous. If there is ever a point where we cannot reasonably assert that the record continues to carry its original message intact, we can never thereafter assert that it is authentic. It is important to recognize that while the process must be continuous over time, the activities that constitute the process are discrete steps. Each instance where the way the information is represented changes – whether moving between storage and use or between storage media or subsystems – is a potential point of failure, a weak link where the entire chain could be broken.' InterPARES

Preservation Task Force Final Report, Draft October 2001, p.88, http://www.interpares.org/documents/ptf_draft_final_report.pdf

[24] DLM Forum, *Guidelines on best practices for using electronic information'*, INSAR supplement III, Office for Official Publications of the European Commission, Luxembourg, 1997, p. 13. Cp InterPARES: 'With electronic records, concern is often expressed about preserving the 'look and feel,' that is the presentation features, of the record; however, there are elements of extrinsic form that the writer cannot fix in an immutable form, but may be changed at whim by any user.' InterPARES *Preservation Task Force Final Report*, Draft October 2001, p.92, http://www.interpares.org/documents/ptf_draft_final_report.pdf

[25] See Jeff Rothenberg and Tora Bikson *Carrying Authentic, Understandable and Usable Digital Records Through Time*. Report to the Dutch National Archives and Ministry of the Interior, 1999, pp. 44-46.

[26] Kenneth Thibodeau "Building the archives of the future'. *D-Lib Magazine*, Volume 7 Number 2 February 2001. http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html Cp also: 'Preservers should assume that future users would want to use the best available technology for access to the records. The design of preservation systems should take into consideration the need to be able to interface with evolving technologies for information discovery, retrieval, communication and presentation.' InterPARES *Preservation Task Force Final Report*, Draft October 2001, p.19, http://www.interpares.org/documents/ptf_draft_final_report.pdf

[27] See Howard Besser, 'Longevity of electronic art', February 2001, http://www.gseis.ucla.edu/~howard/Papers/elect-art-longevity.html and the Guggenheim Variable Media Initiative (http://www.guggenheim.org/variablemedia), a proactive program that asks artists to provide guidelines for presenting their works in new environments.

[28] The model that is now widely adopted to gain a better understanding of which elements and processes are needed for preserving any kind of digital information, the Open Archival Information System (OAIS), distinguishes a Data Object (the bit stream) and Representation Information (which enables the interpretation of the bit stream into meaningful information). See OCLC/RLG Working Group on Preservation Metadata 'A recommendation for content information', report, October 2001, p.3. http://www.oclc.org/research/pmwg/contentinformation.pdf

[29] Several organizations have now published recommendations for preservation metadata for digital resources.;see for instance the CEDARS project http://www.leeds.ac.uk/cedars/OutlineSpec.htm and the final report of the RLG Working Group on Preservation Issues of Metadata http://www.rlg.org./preserv/presmeta.html

[30] Jeff Rothenberg and Tora Bikson *Carrying Authentic, Understandable and Usable Digital Records Through Time*. Report to the Dutch National Archives and Ministry of the Interior, 1999, p.6.

[31] Often a distinction is made between *'born digital' materials* -i.e. that were originally created in digital form- and *digitized materials* -i.e. that have been created by converting the primary, analogue resource into digital files. To facilitate access, cultural institutions are building huge collections by digitizing printed books and journals, photographs and objects.

It is sometimes suggested that preservation of such digitized materials is somehow different, because they have analogue equivalents to fall back on that are preserved in any case.

However, such a distinction does not work for two reasons. The first is that a digitized collection presented on the web usually combines materials of both types, e.g. digital images of photographs that have analogue equivalents, as well as descriptions taken from a catalogue existing only in digital format, as texts, search facilities and lay-out created specifically for the site.The whole is then in fact a new, original product which may have to be preserved in its own right. Second, given the investments made in creating digital collections, it would be a waste of valuable resources not to maintain them properly. The simple fact that there are analogue equivalents that are preserved anyway is no argument for ignoring preservation requirements of digital materials. In defining approaches to digital preservation, the distinction between born digital and digitized materials is therefore more confusing than helpful.

[32] The Public Record Office recommends checks on the stability of media at no more than 5 year intervals. *Management, Appraisal and Preservation of Electronic Records*, Volume 2: Procedures, Chapter 5 : Preservation of electronic records, 5.31. http://www.pro.gov.uk/recordsmanagement/eros/guidelines/procedures5.htm

[33] Gregory W. Lawrence et al, *Risk Management of Digital Information: a File Format Investigation,* Council on Library and Information Resources, 2000, http://www.clir.org/pubs/abstract/pub93abst.html

[34] For an overview of different approaches and the (dis)advantages of each, see Maggie Jones and Neil Beagrie, *Preservation Management of Digital Materials*, British Library, 2001, pp. 102-110.

[35] Gregory W. Lawrence et al, *Risk Management of Digital Information: a File Format Investigation,* Council on Library and Information Resources, 2000, pp. 14-15, http://www.clir.org/pubs/abstract/pub93abst.html

[36] Jeff Rotherbeng, 'Ensuring the longevity of digital documents', *Scientific American,* Volume 272:1, January 1995, pp.42-47.

[37] Donald Waters and John Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access/Rsearch Libraries Group, 1996, p.37. http://www.rlg.org/ArchTF

[38] Maggie Jones and Neil Beagrie, *Preservation Management of Digital Materials. A Handbook*, British Library, 2001, p.58.

[39] As the Public Record Office states: 'Departmental preservation strategies must provide for long-term preservation; that is for periods of five years or longer' (*Management, Appraisal and Preservation of Electronic Records*, Volume 2: Procedures, Chapter 5 : Preservation of electronic records, 5.14.http://www.pro.gov.uk/recordsmanagement/eros/guidelines/procedures5.htm). In the old world, it would have been unthinkable to refer to a period of five years as long term.

[40] See http://www.elsevier.com/inca/publications/misc/ni2164.pdf

[41] Message posted on IFLA-L January 17 2002.

[42] The Koninklijke Bibliotheek, the national library of the Netherlands, and the Dutch Publishers Association have made such an experimental agreement specifying storage, copying, access etc. See http://www.kb.nl/kb/dnp/overeenkomst-nuv-kb-en.pdf

[43] *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources,* RLG-OCLC Report, Draft, August 2001, pp. 19-20, http://www.rlg.org./longterm/attribswg.html.

[44] Donald Waters and John Garrett, *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information*, Commission on Preservation and Access/Research Libraries Group, 1996, p.37. http://www.rlg.org/ArchTF

[45] See http://www.kb.nl/coop/nedlib/index.html

[46] See http://ssdoo.gsfc.nasa.gov/nost/isoas/ref_model.html

[47] A detailed discussion, based on the Open Archival Information System as a reference model, is presented in *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources,* RLG-OCLC Report, Draft, August 2001, http://www.rlg.org./longterm/attribswg.html.