ARCHIVING CHALLENGES AND TECHNOLOGICAL APPROACHES

NACLIN-2009

S. Venkadesan,

JRD Tata Memorial Library, Indian Institute of Science, Bangalore-560012.

venky@library.iisc.ernet.in

Introduction

- In recent decades, the core processes of organizations, companies, governments and individuals have become increasingly dependent on digital objects of all kinds—including documents, images, videos, sound files, spreadsheets, datasets (such as patient records), executable programs and integrated combinations thereof.
- But the pace of technological change and the nature of digital objects make them obsolete and unusable in just a few years. In particular, 'inherently digital' artefacts that are not simply page images—and are characterised by complex dynamic, executable and interactive behaviour— are likely to become increasingly important over time.
- If preservation methods cannot perpetuate their full range of behaviour, the future scholarly record will at best bear only static, snapshot representations of the first generation of inherently digital objects; at worst it may bear no meaningful trace of them at all.
- At a time when digital objects are being generated everywhere and technology is changing at unprecedented speed, developing an archiving strategy for these objects is like chasing a moving train.

Introduction

- Today's researchers need to be able to 'stand on the shoulders of giants' through access to the work of the existing body of knowledge. Academic progress is marked by articles published in peer-reviewed journals. Tens of thousands of scholarly journals published worldwide distribute an estimated 1.4 million articles a year.
- Libraries and other institutions have traditionally archived these journals, building massive physical collections that provide researchers with originals or facsimiles (such as microfilm or photocopies) of published work. Paper and photographic page images have so far proved a durable and accurate way of obtaining access to the past.
- But it is uncertain whether this traditional preservation approach will be sufficiently robust for the ways in which research results will be circulated, reviewed, accessed and maintained in the future.
- Archiving and preserving digital objects is fundamentally different from archiving and preserving print objects. Digital objects must be rendered into human-readable form by executable software, and they may become unreadable in just a few years because their formats have become obsolete.
- Preserving these objects for the long term in a form that ensures future access to their original intellectual content requires a substantial investment in infrastructure, equipment, skills and expertise.

Why Archive? – the Key driving forces

25-Sep-09

- Many regard the archiving and preservation business model—who provides access to whom, in what form, at what time, and who pays—as the biggest challenge for publishers, libraries and other stakeholders.
- This business model is beset with uncertainties that make the development of archiving strategies extremely difficult.
- However, behind these uncertainties one finds a number of identifiable forces driving the future of digital preservation.
- Four of these are discussed.

Why Archive? – the Key driving forces

25-Sep-09

How will research be communicated in the future?

- Scholarly communication is fragmenting as it evolves from the traditional model of publication in peer-reviewed journals to wider access through electronic media and information hubs.
- A number of factors influence this, not least the cost of subscriptions to journals and the fact that paper-based publications cannot support the active nature and pace of new forms of academic research.
- But while alternative dissemination channels are emerging, existing tenure and promotion mechanisms are still an incentive for scholars to publish papers in traditional peer-reviewed journals.
- Additionally, published articles are developing an increasingly 'long tail' of economic value, spurred by new digital opportunities for repackaging, reselling and improving access.

Why Archive? – the Key driving forces

25-Sep-09

Who preserves what?

- As journals are increasingly sold in the form of licences that give access to online content, libraries are no longer in control of their archives.
- When a journal ceases publication or its publisher goes out of business, its online content may be lost if the long-tail value of the articles is not commercially attractive.
- Traditionally, national libraries have stepped in to guarantee that access to national research output is guaranteed in perpetuity.
- However, with the increasing internationalization of research and the vanishing boundaries of the digital environment, international initiatives will be required.

Why Archive? – the Key driving forces

25-Sep-09

Who pays for what?

- Stakeholders in digital objects have different needs and perspectives, and the optimum funding model for preservation has yet to be determined.
- One obvious answer might be government funding, but a private approach may be equally—or even more—effective.
- Archiving and preservation may be regarded as an insurance policy for guaranteed access to digital content, but who will be willing to pay the premium?

Why Archive? – the Key driving forces

What do research libraries and universities demand of archiving and preservation services?

- As yet there is not wide concern in the scholarly community about preserving the original behavior of digital artifacts; the focus is mainly on guaranteeing access to future 'vernacular' versions of page-image documents.
- Most current digital archiving initiatives do not therefore attempt to preserve the full functionality of digital objects.
- However, that functionality may be a crucial part of a digital object's intellectual content. Furthermore, historians, the general public, governments and nonprofit institutions may require the preservation of such functionality, whether to ensure intellectual validity, recreate historical perspective, enable aesthetic appreciation, enforce legal and ethical accountability or simply venerate important artifacts. For all of these reasons it may be necessary to retain the original behavior (as well as the look and feel) of digital objects, particularly those that are inherently digital.

Archiving Challenges

- Electronic archiving of scholarly journals is an important issue for libraries. This issue cannot be neglected any more because the usage of electronic journals has increased significantly in recent years.
- With the greater capability of the digital medium, however, the content of digital files may be lost to future scholars not just because the physical item deteriorates, but because the information cannot be extracted and interpreted correctly.
- A scholarly journal on the printed page can be viewed and read without any special equipment as long as one knows the language in which it is written.
- Digital scholarly journals, however, cannot be viewed without special equipment, such as a computer, an Internet connection, and the required software.
- Unlike paper or microfilm where the meaning is transparently inscribed on the surface of the medium— digital documents are opaque bit streams only understandable to humans when interpreted by a machine.
- The hardware and software to do this interpretation, however, is constantly superseded. There have, for instance, been more than 200 digital storage formats alone deployed since the 1960s, with none lasting more than 10 years.

Archiving Challenges

- With the machine dependency, archiving of electronic journals is more complicated than archiving print journals.
- The short lifecycle of digital media is a threat for digital archiving because digital media become obsolete much faster than print media. The format of the digital resources can be damaged or lost and may no longer be intact, retrievable, understandable, or displayable. The technology used to store the publication is likely to become obsolete even before that happens. Therefore, continued access to archived resources is a big issue in digital archiving, while "access" was not a big issue to traditional archiving.
- Developments in information technology have also changed the traditional system of publishing, distributing, and even the use of scholarly journals. The initial communication for publishing a paper is so quick now, especially with the help of e-mail. An accepted manuscript can be accessed online before the date of publication. The distribution cost of journals has also decreased in the digital environment. Even the patterns of use of scholarly journals are changing in the digital environment.
- Moreover, information technology has caused substantial changes in the traditional roles of libraries and publishers. One of the major changes is a shift in responsibility of archiving from libraries to publishers in an electronic environment.

Archiving Challenges

- Historically, archiving records and documents has long been the responsibility of librarians, and publishers largely shied away from this role. Several libraries such as the IISc library, hold many research journals in print from their first volumes. Few publishers have complete journal collections archived for posterity.
- Regarding legal issues, the Internet and electronic journals make the task of illegal copying and distribution easier. Equally, the Internet and electronic journals make the task of policing a very difficult one. While copyright permits only fair use and prohibits all distribution whether for profit or non-profit, enforcing copyright in the electronic environment is a complex task for publishers, institutions, and governments. Internet and electronic journals have created new paradigms for archiving journals, resulting in controversial debates, especially in light of new technological complexities.
- In the shift from print to electronic formats, the story of ownership is also changing. One reason is when publishers arrange for scholarly journals online through licensing agreements, libraries do not have local possession of a copy as they did with print nor do they own the publications. As a consequence, the archiving of printed publications implicitly offered by library collections no longer exists for electronic journals. If a publisher fails to maintain its archive for any reason, there would not be any access to those resources.

Archiving Challenges

- One of the problems with archiving electronic journals is the mergers of publishers. Several significant mergers have occurred in recent years. What happened to back issues of these journals? This is a key question that publishers must consider. For Academic Press, all of its titles were migrated from their platform to Science-Direct. Academic Press sent their journals to the KB (Koninklijke Bibliotheek or the National Library of the Netherlands) at the time they were added to Science-Direct. Now, all of the Academic Press titles and their backfiles are at the KB.
- In migrating from print to the electronic environment, standards play an important role in electronic archiving; however, there is a debate over the success of standards for computers.
- According to the National Library of Australia's initiative for preserving digital resources (PADI), "resources which are encoded using open standards have a greater chance of remaining accessible after an extended period than resources encoded with proprietary standards."
- At present, technical issues and challenges related to digital preservation include a lack of practical implementations of preservation standards and a lack of technical knowledge, in general, of what information is required to support the digital preservation process within the institutions.

Other Archiving Considerations

- "Selection" is another important issue in electronic archiving. The huge quantity of information being produced digitally, its variable quality, and the resource constraints on those taking responsibility to preserve long-term access make selectivity inevitable for archiving.
- Traditionally, lack of selection for preservation may not necessarily mean that the item will be lost, but in the digital environment non-selection for preservation will almost certainly mean loss of the item.
- Although not all resources can or need to be preserved forever, some will not need to be preserved at all, others will need to be preserved only for a defined period of time, and a relatively small sub-set will need to be preserved indefinitely.
- In traditional archiving, some level of redundancy with multiple copies was inevitable in different repositories, but the story is different in the electronic environment. There was large-scale redundancy in the storage of journals in the print era, as many different institutions collected the same titles. Theoretically, in a digital environment, a single institution can provide worldwide access and accept preservation responsibility, although there is a debate whether a level of redundancy should exist in the digital environment.

Other Archiving Considerations

- Copyright seems to be established well for traditional archiving but not for electronic materials. The copyright and intellectual property rights issues for digital materials are much more complex and significant than for traditional media. If these issues are not addressed adequately, preservation will be curtailed.
- Both publishers and librarians perceive threats in the digital environment. Some authors are quite happy for their material to be widely accessible, while some publishers do feel threatened.
- The ease of distribution and duplication offered by new technologies raises commercial concerns and has driven publishers to seek control of content through legal and technical tools, such as licensing and digital rights management.
- But libraries are involved in providing access to information, and librarians have concerns that new approaches threaten legitimate access to the detriment of the public good. Although making changes in law or licensing practice is difficult, rights holders and libraries have to understand and cooperate with each other to progress.

Other Archiving Considerations

- There are many stakeholders such as authors, publishers, libraries, archive centers, distributors, networked information service providers, IT suppliers, legal depositories, consortia, universities, and research funders, who may have an interest in archiving electronic journals. It is important to consider the relationship of the stakeholder to the digital material archiving.

Cooperation and Communication

- Electronic archiving is expensive and creates new levels of responsibility for publishers and for libraries, involving functions similar to data storage. Electronic archiving requires considerable costs for infrastructure and maintenance.
- If the publishers can develop an institutional model and legal framework to guarantee perpetual access to the subscribed journals for every subscriber, libraries may need not archives.
- However, the geo-political factors, the frequent corporate changes at publishing organizations in terms of mergers, and the changes in the structure and ownership of journal become compelling factors for the library to think of independent local archiving.
- Apart from the high cost of infrastructure and maintenance, archiving will be a substantially repetitive cost for every library in a community. Moreover, it can be a cooperative task and cost-sharing activity among the libraries in a community who can assign the responsibility to one of its members.
- Shared responsibility, however, would not be enough to secure archiving of electronic journals.
- Libraries and archives already realize the importance of archiving and preserving continued access to digital materials, and many institutions have begun to take initial steps to meet their responsibility effectively.

Cooperation and Communication

- The new concerns of electronic archiving have led to a series of meetings over the past few years among publishers, librarians, and technologists and sponsored by a variety of organizations. A number of the academic librarians, university administrators from the United States, and others participated in a meeting to discuss electronic journal preservation at the Andrew W. Mellon Foundation offices in New York on September 13, 2005. They suggested four key essential actions regarding archiving :

  - First, research and academic libraries and associated academic institutions must recognize that preservation of electronic journals is a kind of insurance, and is not in and of itself a form of access.
  - Second, in order to address these risk factors and provide insurance against loss, qualified preservation archives would provide a minimal set of well-defined services.
  - Third, libraries must invest in a qualified archiving solution.
  - Finally, research and academic libraries and associated academic institutions must effectively demand archival deposit by publishers as a condition of licensing electronic journals.

Cooperation and Communication

- They pointed out that universities, colleges, and their libraries have recently been working together to help scholars manage their copyrights and to craft alternatives to high-priced forms of scholarly publishing. It is now equally important that research and academic libraries work with scholars – and their publishers – to sustain future research and teaching by establishing trusted archives in which the published scholarly record in electronic form can persist outside of the exclusive control of publishers, and in the control of entities that value long-term persistence.
- The long-term archiving of electronic journals is one of the important challenges for libraries and publishers. While libraries and publishers have had stable roles for centuries (publishers produced information; libraries provided access to this information), the evolution of the information technology has disrupted this critical role for libraries.
- Librarians currently face many issues and concerns for archiving electronic journals, such as differences between digital and print media, rapid obsolescence of digital technology, shift in the responsibility of archiving to publishers, legal issues such as copyright and intellectual property rights, selection, and many more.
- On the other side, scientific publishers have become aware of the issues and are rising to the challenge by implementing long-term preservation policies.

25-Sep-09

Archiving policies of publishers

- One study focused on the archiving policies of few leading publishers.
- This study showed that while commercial publishers are dominant in the publishing of scholarly electronic journals, they have also taken the responsibility of archiving more seriously than not-for-profit publishers. They have systematically tried to have a long-term preservation for their journals.
- The National Library of the Netherlands or Koninklijke Bibliotheek (KB) has become the main library which several for-profit publishers made an agreement with. After the agreement with Elsevier, who was the first, the KB concluded similar agreements with Kluwer Academic Publishers (2003), BioMed Central (2003), Blackwell (2004), Oxford University Press (2004), Taylor & Francis (2004), Sage (2005), Springer (2005), and Brill Academic Publishers (2005).
- All KB agreements dictate that the KB will preserve what the publisher sends to the library. The archived content is exactly the same as the published content. This coverage may change as publications become more complex and include multimedia and dynamic content. For now, however, the KB's policy is to preserve "as is."
- As a part of the agreements, the KB provides on-site access to the journals on a current basis to all on-site, authorized library users. The agreement covers new publications, as well as digitized backfiles.

Archiving policies of publishers

- In addition, should there be a catastrophic disaster such that the publisher is inoperable for a long period of time, the KB would be part of the interim service system.
- The agreements between the KB and leading publishers of scholarly electronic journals are significant developments in keeping digital archives available in perpetuity. The relationship evokes the traditional role of the library, particularly of national libraries, in undertaking preservation responsibility, while also asserting the commercial role of the publisher.
- Among the top not-for-profit publishers, only Oxford University Press made an agreement with the KB. The University of Cambridge is working with the Massachusetts Institute of Technology (MIT) Libraries to establish a digital repository based on the DSpace software.
- IEEE is following its own archiving policy and prepared the electronic material on CDs. Though the number of electronic journals published by IEEE and not-for-profit publishers in general is less than for-profit publishers, CD-ROMs are not be considered adequate storage for long-term preservation. The CD-ROMs are digital media with a short lifecycle; therefore, IEEE needs to reconsider its archiving policy.

Archiving policies of publishers

- Despite these agreements and developments, the field of digital archiving is still in its infancy, and much work needs to be accomplished to achieve a secure and permanent archiving of electronic journals.
- However, the successful agreements between the KB and the leading for-profit publishers could be used as a model for other publishers, especially for not-for-profit publishers and other publishers around the world.
- Like the KB initiative, there are many other related services such as
  - PORTICO,
  - LOCKSS (Lots of Copies Keep Stuff Safe) and
  - CLOCKSS (Controlled LOCKSS).

Technological Approaches to Archiving

- We not only consider valuable original manuscripts, sound and image carriers as worth preserving, but also all presently available book inventory titles, as well as the rapidly increasing amount of digital data media.
- Whereas ancient tomes require expensive preservation measures and works printed on acidic paper must be safeguarded from total destruction by storing them on microfilm, digital media seem to be characterized by ideal properties: bit sequences can be kept in principle for long periods without any loss of information. The continuously rising storage capacity already permits the storage of extensive reference books on one CD.
- With progressing miniaturization of the data media, it seems that digital storage has become an attractive, universally applicable solution for archiving, which still has further advantages: searches can be performed not only among the entries in electronic catalogs, but also, and within seconds, throughout the entire contents of digital documents.
- Just as rapidly, the Internet brings documents from the whole world to our workplace or home, where we can process them with various tools to merge them into new documents.

Technological Approaches to Archiving

So where is the problem?

- Whoever has used computers as text processors for a long time, knows several answers: first of all, digital data media are not directly readable by humans. Although we are able to perceive cave paintings, which are thousands of years old, in the same manner and with the same sense organs as the contemporaries of the artist, the contents of the text files produced by us become legible for us only indirectly with the help of a suitable computer system. In itself that is not a problem.
- Similarly, the problem that storage media such as floppy diskettes, hard disks, and CDs can age and fail can be met with suitable measures (backup copies) in such a way that no data loss occurs.
- However, the components of the computer systems age even faster than the data media. Nowadays, it is hardly possible to find suitable floppy drives able to read 5.25- or 8-in. floppy diskettes, and even if such drives were to be found, it would not be possible to connect them to modern computers due to the lack of suitably updated driver software.

Technological Approaches to Archiving

- Concerning software, variety and change is even more important.
- Over the years, thousands of different editors for digital documents (text, graphics, spread-sheet analysis, slide, Web page editors, etc.) have been developed, most of which having proprietary internal document data formats.
- Due to the incompatibility of these formats, it is often impossible to freely exchange documents among different editors. At least, such exchanges will likely entail substantial quality losses. In the long run, only comparatively few of these systems will withstand marketplace competition. Users of commercially less successful products sooner or later are confronted with the problem of finding a way to save and transfer their collection of outdated documents to a new software environment.
- This problem is usually solved by means of labor-intensive “heroic measures” like recreating the most important documents from scratch or scanning printed texts into new systems, with subsequent manual correction work. Obviously, this procedure is not suitable for the enormous document collections of libraries and archives, in which despite limited manpower both collection preservation and document access must be guaranteed for an indefinite period of time.

Technological Approaches to Archiving

- Data are stored in computers as sequences of ones and zeros, i.e., so called bit-streams. Essentially, digital documents are represented by bit streams.
- In more general terms, they are digital character streams, since other basic character sets can be used instead of just zeros and ones. The essential problem to long term preservation of digital objects still to be solved is how to interpret the enormous variety of multimedia formats which give "meaning" to the conserved binary or text-based data streams.
- After long periods of time, the textual, graphic, and acoustic components of multimedia documents have to be recreated and rendered to produce the same effects that were originally perceived by their creators.
- For interactive documents like hypertext, animations, or spreadsheets, we additionally require that reactions to user inputs be true to the original.
- It's not just documents in libraries that may need to be archived —examples of other digital objects that may be endangered include: Spreadsheets, Patient records, Meteorological data, Seismological data, Multimedia works, Models and simulations, Geographical Information Systems data and the Web content.

Migration

- In IT a Migration approach is employed almost universally, wherein documents are continually "migrated" from one data format to another. Using migration for preservation of digital documents has some obvious advantages. There is a rich body of knowledge about methods, plenty of trained staff, and, in particular, a set of available tools.
- In principle, migrated documents are available on some current rendition system any time. In the absence of legal (or other) obstacles, even world-wide access should be no problem.
- Migrated documents typically satisfy current quality standards: In the same way stereophonic and hi-fi quality can be taken for granted with current audio recordings, a modern text document can be expected to contain up-to-date fonts and layouts.
- Migration also often requires or entails small adjustments of the original document in order to make the result fit into the new environment. In a number of migration steps these ever so small adaptations sum up and thus inevitably jeopardize the authenticity of the resulting document.
- Only simple migration steps can be fully automated, i.e., performed by a program. An instance of such a comparatively simple migration step is the conversion of ASCII documents into the Unicode format. Even here, the result needs to be checked manually for places where authors of ASCII documents have used circumscriptions for symbols that can now be represented directly in Unicode.

Hardware Museums

- The approach, which is most directly oriented towards keeping up the original state, is to establish so-called hardware museums.
- The mission of a hardware museum is to collect (and keep operational) all relevant computing systems so that future generations may view our documents in their original environments.
- There are a few hardware museum for eg. One at the Universität der Bundeswehr München, Germany *or* **Computer History Museum,** Mountain View, CA, for accessing outdated data media.
- The authenticity of this approach obviously cannot be surpassed.








Hardware Museums

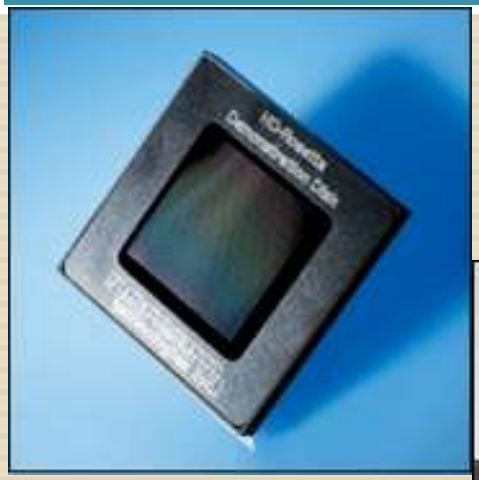
- Unfortunately, it is not a feasible solution in practice for a number of reasons:
  - The endeavor to assemble in hardware museums all computers that have ever been developed has not been carried out so far. With every new generation of computers, it becomes more unrealistic because of the enormous number of items to be collected.
  - A complete rendition system consists of hardware, systems software, and those application programs that actually render the documents.
  - Hence, in addition to the hardware, a hardware museum would have to collect all the different versions of systems and application software (and to install any combination of systems and application software as required).
  - This requirement extends to additional hardware devices (like joysticks, special graphics cards, plus their driver software).
  - The durability of technical devices and their components is limited. Since outdated devices and components are not produced any more, one would have to archive precise technical specifications of both the complete computer architecture and of all its components in order to give future maintenance engineers a chance to remove defects and to reproduce components that have to be replaced (always hoping that the techniques for producing such devices are still understood reasonably well).
  - Experts agree that a hardware museums project would be doomed to fail because of the enormous, ever increasing costs entailed.

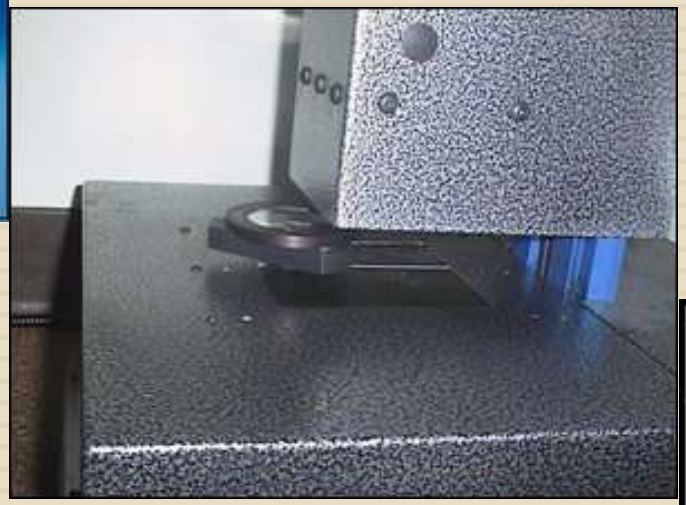
Analogue Preservation on Silicon Wafers

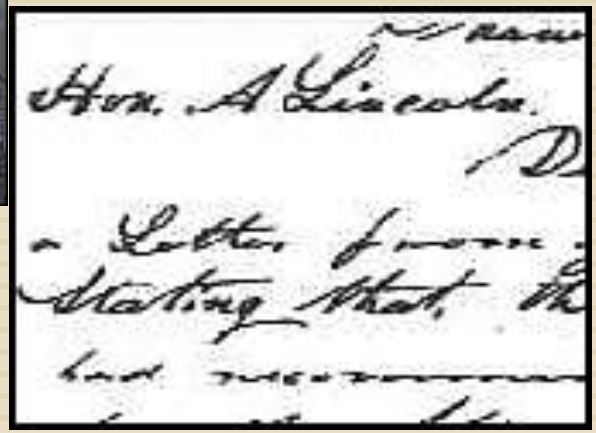
- High Density Rosetta (HD-Rosetta) provides analog storage of information and images on nickel plates that last for thousands of years. HD-Rosetta is patented worldwide by Norsam under exclusive license from Los Alamos National Laboratory. Each disc is 2.2 inches in diameter and contains approximately 9000 pages of text or images. The discs are made by covering a silicon wafer substrate with a thin (~400 micrometer) layer of nickel. The nickel layer, which contains the text, is formed from a master disc using an electroforming process. The height of the lines that make up the letters and images on the master disc is very small, less than about 100 nm.
- **The HD-Rosetta writing process:** Each page is digitally converted, or digital files are received, minimized to microns and written onto the nickel plate pixel by pixel, using a Focused Ion Beam (FIB) machine. Inside the FIB, gallium ions are forced down through the machine's main column onto the surface of the plate. The gallium ions essentially knock off atoms from the surface and micro-engrave into any given medium.
- **The HD-Rosetta reading process:** Depending upon how many pages are stored on the Rosetta, users may employ viewers as simple as optical microscopes to read and retrieve the etched information, or for higher densities, electron microscopes are employed. Norsam is developing a special HD-Rosetta reader which locates x,y,z coordinates.

Analogue Preservation on Silicon Wafers

25-Sep-09







Analogue Preservation on Silicon Wafers

25-Sep-09

Advantages of the Rosetta:

- ***Compact :*** On a plate no larger than 2" square and 1/4" thick, the Rosetta has immense storage capabilities. You may store about 196,000 pages with electron microscope retrieval between 5,000 - 18,000 pages with optical microscope retrieval.
- ***Technologically Secure:*** Unlike digital data, which require medium-specific software and operating systems to retrieve information, the Rosetta needs only a simple magnifying lens to view the analog data. It will never face obsolescence.
- ***Magnetically Stable*** :Since the Rosetta's information is etched into nickel, it is never affected by electromagnetic radiations as is often the case with digital data.
- ***Physically Durable*** :With its magnetic immunity, a life expectancy of at least 1,000 years and a temperature threshold of 500° C, nickel is the HD-Rosetta material of choice.
- ***Easily Generated*** :Each page takes less than 1/10th of a second to write. In about two hours, our automated machines can write approximately 7,000 pages.
- ***Permanently Readable*** :Because the HD-Rosetta information is stored in analog format, the data can be written in any language (images included), and it can be read by anyone with a simple magnifier for as long as the plate endures. And because of the Rosetta's resilient nickel structure, this plate will endure indefinitely.

Emulation

- In order to save time and costs, so-called emulators are employed for the development of new hardware.
- Emulators allow the function of processors and other hardware components to be simulated by software. Thus hardware can be tested thoroughly before being actually built.

- When using emulation, for each digital document the following items have to be preserved (using, e.g., migration):
  - The character stream and the metadata
  - A specification of the hardware that can be interpreted by the emulator
  - The complete software of the rendition system (in the form of binary data streams).

Emulation

To access a document conserved through Emulation, say, 100 years from now, one would have to proceed as follows:

- Load the hardware specification into an emulator to obtain a software implementation which is functionally equivalent to the original hardware. This is the most important step and requires the format of the hardware specification to be a suitable input for the emulator software used.
- On the emulated computer install the systems software and the application programs needed for rendering the document. For using this 100-years-old computer, in particular for installing software, we must have the manuals. This is why the manuals have to be archived as part of the metadata.
- Load the character stream of the digital document into the emulated computer and start the rendition software to access the document. For this purpose, the manuals for the rendition software must be available and, therefore, should be archived as part of the metadata, too.

Emulation

In spite of its evident complexity, using emulation for long-term preservation of digital documents is attractive for a number of reasons:

- Once an digital document has been prepared for long-time preservation via emulation – recall that along with its character stream the emulator specification, the software of the rendition system, and all the metadata have to be stored – only regular refreshing of these items is required in order to preserve the document for an unbounded period of time.
- For accessing documents of a certain type the rather costly steps 1 and 2 above have to be performed, but only rarely. Like with the hardware museum approach, each document is represented by its original binary data stream.
- All the other information required for preservation via emulation (emulator specification, software, metadata) has to be prepared and stored only once for each computer hardware. Most of the preparation effort goes into the hardware specification. Therefore, it would be desirable to directly obtain such specifications from the hardware manufacturers (who are using emulators for hardware development anyhow).
- Since with this approach the original data stream is never affected by any changes, its authenticity is very high, always assuming precise hardware specifications.
- Whether emulation is a universally applicable way of preserving digital documents, is still an open question.

Standard Format

- The costs and effort involved in long-term preservation are proportional to the number of data formats involved. This is particularly true for migration, but also holds for emulation. An obvious solution is to support only a restricted set of formats which includes just enough formats as to satisfy all sensible purposes, but not all variations thereof.
- The standards for simple character sequences (ASCII, Unicode), are well established and should be stable for some time to come. Hence, they are particularly well suited for storing metadata, which are the key to the contents and the handling of documents.
- For complex document types that are often used, we typically have a great variety of different, rivaling formats. Considering, e.g., text documents with embedded graphics, the spectrum ranges from internal, proprietary formats (say, MS Word), procedural descriptions (PostScript), and viewing formats (DVI and PDF) to markup languages (LATEX and XSL).
- Currently, on the basis of XML a number of "self-describing formats" is emerging that will probably give direction to future standardization efforts. These include the metadata standards RDF and Dublin Core, XSL for desktop publishing, SVG for vector graphics, SMIL for multimedia documents, and MathML for mathematical formulae.

Other Legal and Social Concerns

- Before an digital document may actually be filed into an archive, copyright questions have to be resolved: migration demands documents to be copied and often requires minor modifications. The authors of the original document must be asked in advance to permit such actions.
- Readers have to be charged appropriately for printed copies. For digital copies, readers must additionally asked to sign license agreements for documents that are not freely available.
- With the emulation approach we also have to consider (at least in the medium term) reserved rights concerning the different software components of the rendition environment. Again, some fixed-rate royalty agreements for permissible purpose access would be a great convenience.
- If the emulation approach is to be used on a large scale, ideally hardware manufacturers should be bound by law to publish complete specifications of their hardware in some suitable emulator input format.
- Since currently available hardware needs not to be emulated, such an obligation would only have to take effect as soon as the hardware is not produced any more, i.e., becomes obsolete. At this point, the producers will have drawn the full economic benefit from their short-lived innovations.

Other Legal and Social Concerns

- Libraries and archives need precise criteria on what is to be considered valuable and, therefore, should be conserved.
- For print media there are formal criteria (e.g., whether there is an associated ISBN number) and policies supported by law (publishers have to provide specimen copies). In order to do full justice to the relevance of digital documents, we also have to consider documents that are not available on tangible data carriers like CD-ROMs, and that do not have an ISBN associated with them.
- Scientific on-line journals and other publications that are "only" available from the Internet, certainly also belong to the cultural heritage, which is worth to be preserved.
- For the new media, new criteria and new policies have to be established. Because of the wealth of available materials and because of the high costs involved we need carefully elaborated selection criteria.
- Costs are an all-dominant limiting factor. Both on the national and international levels we need to arrive at sustainable long-term agreements about who are the bodies that are legally and economically responsible for the long-term preservation of digital documents, and what competencies and budgets they will be given.
- It also makes sense to reduce costs by cooperation and division of functions. By standardizing input formats for both documents and metadata, the archiving costs can be further reduced.
- Expenses can be either borne exclusively by the state. Alternatively, some of the costs can be accepted by customers, e.g., by paying fees for lending and other services. Authors who want their works to be preserved might well be willing to contribute.
- Regulations should balance the legitimate interests of all share holders- authors and publishing companies, libraries and archives, scientists, and the general public.

Recommended Reading

I. Long-Term Preservation of Digital Documents: Principles and Practices
Borghoff, U.M., Rödig, P., Scheffczyk, J., Schmitz, L. 2006, Springer
ISBN: 978-3-540-33639-6

II. Digital preservation: The uncertain future of saving the past
Stijn Hoorens, Jeff Rothenberg, RAND Corporation research brief series
Availability: Web-Only Pages: 4, Document Number: RB-9331-RE, Year: 2008
http://www.rand.org/pubs/research_briefs/RB9331/

III. Archiving challenges of scholarly electronic journals: How do publishers manage them? Moghaddam, G. G., Serials Review, 33(2), 81-90. (2007).

Thank You!

Questions