Digital Preservation

Preservation Strategies

by

Dr. Jagdish Arora
Director, INFLIBNET Centre
jarora@inflibnet.ac.in

Digital Preservation Strategies

- **Short-term Strategies**

  - **Bit-stream Copying**

  - **Refreshing**

  - **Replication**

  - **Technology Preservation or Computer Museum**

  - **Backwards Compatibility and Version Migration**

Digital Preservation Strategies

- **Medium- to Long-term Strategies**

  - **Migration**
  - **Viewers and Migration at the Point of Access**
  - **Canonicalization**
  - **Emulation**

- **Investment Strategies**

  - **Restricting Range of Formats and Standards**
  - **Reliance on Standards**
  - **Data Abstraction and Structuring**
  - **Encapsulation**
  - **Software Re-engineering**
  - **Universal Virtual Computer**

Digital Preservation Strategies

◆ **Alternative strategies**

　◆ **Analogue Backups**
　◆ **Digital Archaeology or Data Recovery**

◆ **Combinations**

Digital Preservation Strategies

Short-term Strategies

Short-term Strategies

Bit-stream Copying

- **Known as "backing up data"; making an exact duplicate of a digital object.**

- **Remote storage guard against disastrous event.**

- **Minimum maintenance strategy for even the most lightly valued, ephemeral data.**

- **Not a long-term preservation strategy.**

Refreshing

- copying digital information from one long-term storage medium to another with no change in the bit-stream

- Addresses both decay and obsolescence issues related to the storage media.

- Does not address the issue of obsolescence of encoding and formatting schemes.

- Longitivity of media does not guarantee availability of hardware / software required to read the stored format.

- Backward compatibility and interoperability are serious threat to longevity of digital information.

Replication

- **Replication is used to represent multiple digital preservation strategies.**

- **Bit-stream copying is a form of replication.**

- **LOCKSS (Lots of Copies Keeps Stuff Safe) is a consortial form of replication, while peer-to-peer data trading is an open, free-market form of replication.**

- **Objective is to enhance the longevity of digital documents while maintaining their authenticity and integrity through copying and the use of multiple storage locations.**

Technology Preservation

- Also called the "computer museum" solution.

- Rely on preserving the computer, operating systems, original application software, media drives, etc.

- Applicable for neglected digital objects.

- Assumes that media has not decayed beyond readability.

- Limitation: No obsolete technology can be kept functional indefinitely.

- Requires a considerable investment in equipment and personnel.

Backward Compatibility & Version Migration

- Current versions of software can interpret and present digital material created with previous versions of the same software.

- Option for version migration converts documents into the current format.

- MS Word, Excel and Access applications, allow previous versions of their file formats to be transformed and resaved in a new version

- Option not be available for all types of objects.

- The process of migration is likely to introduce unwanted changes to a document incrementally if used over many generations.

Digital Preservation Strategies

Medium to Long-term Strategies

Migration

- **Periodic transfer of digital materials from one hardware / software configuration to another, or from one generation of computer technology to a subsequent generations.**

- **Migration may include conversion of data to avoid obsolescence not only of the physical storage medium, but of the encoding and format of the data.**

- **Digital objects will have to be constantly migrated and converted to new formats, computing devices, storage media and software to ensure they are not left behind on obsolete system.**

Viewers and and Migration at the Point of Access

- An alternative to recurring migration. Involves use of viewers, software tools that provide accessibility at the time of access, using the original data stream.

- Example: The "migration on request" approach has been developed in CEDARS and CAMiLEON projects that uses a software tool to record method of access.

- Google's Google Docs

- Limitations i) viewers not be available for all formats; ii) Viewers may represent some, but not all; iii) the gap between the original format and the prevailing technologies at the time of access may be too much to tackle

Canonicalization

- **Technique designed to determine maintenance of essential characteristics of a document through conversion from one format to another.**

- **Canonicalization relies on the creation of a representation of a digital object that conveys all its key aspects.**

- **Once created, this form could be used to algorithmically verify that a converted file has not lost any of its essence.**

Emulation

+ **Emulation uses a special type of software, called an emulator, to translate instructions from original software to execute on new platforms.**

+ **Eliminate the need to keep old hardware working.**

+ **Emulation requires the creation of emulator programs that translate code and instructions from one computing environment so it can be properly executed in another.**

Software Re-engineering

- Software reengineering may offer a number of strategies for transforming software as technologies change.

  - Adjustment and re-compiling of source code for a new platform;

  - Re-coding of the software from scratch, or re-coding in another programming language; and

  - Translation of compiled binary instructions for one platform directly into binary instructions for another platform.

- Requires source code. Porting to other platforms requires considerable time and effort.

Universal Virtual Computer

- A form of emulation.

- Requires development of a computer program independent of existing hardware or software that could simulate the basic architecture of every computer.

- Users could create and save digital files using software of their choice, but all files would be backed up in a way that could be read by the universal computer.

- Reading a file in future would require only a single emulation layer—between the universal virtual computer and the computer of that time.

Digital Preservation Strategies

Investment Strategies

Restricting Range of Formats and Standards

- **Preservation programmes may decide to only store data in a limited range of formats and standards.**

- **All digital objects within an archival repository of a particular type can be converted into a single chosen file format that is thought to embody the best overall compromise amongst characteristics such as functionality, longevity, and preservability.**

- **For, example most of the textual and graphical information can be converted into PDF format.**

- **The strategy does not solve the access problem of obsolescence of formats and standards**

Reliance on Standards

- **Advocates use of well-recognized standards and discarding proprietary or less-supported standards.**

- **Backward compatibility for older formats would maintained if it is widely used as a standard.**

- **For example, if JPEG2000 becomes a widely adopted standard, the sheer volume of users will guarantee that software to encode, decode, and render JPEG2000 images will be upgraded to meet the demands of new operating systems, CPUs, etc.**

Data Abstraction and Structuring

- **Data abstraction involves analyzing and tagging data so that the functions, relationships and structure of specific elements can be described.**

- **Using data abstraction, the representation of content can be liberated from specific software applications and be achieved using different applications as technology changes.**

- **The technique requires extensive development of tools and methods for analysis and processing in order to correctly represent and tag each type of data.**

Encapsulation

- Technique of grouping together digital objects and metadata necessary to provide access to that object.

- The grouping process lessens the possibility that any critical component necessary to decode and render a digital object will be lost.

- Appropriate types of metadata to encapsulate with a digital object include reference, representation, provenance, fixity and context information. Encapsulation is considered a key element of emulation.

Digital Preservation Strategies

Alternative Strategies

Digital Archaeology

- Rescue content from damaged media or from damaged hardware and software

- An emergency recovery strategy involves specialized techniques to recover data from unreadable media, either due to physical damage or hardware failure.

- Carried out by data recovery companies

- Given enough resources, readable bit-streams can often be recovered even from heavily damaged media (especially magnetic media)

Analogue Backups

- **Combines the conversion of digital objects into analogue form, e.g., taking high-quality printouts or the creation of microfilm.**

- **An analogue copy of a digital object can, in some respects, preserve its content and protect it from obsolescence.**

- **Technique makes sense for documents whose contents merit the highest level of redundancy and protection from loss.**

Conclusion

No single digital preservation strategy can serve as a practical solution to the problem of technological obsolescence for digital materials.

No single solution that is appropriate for all data types, situations, or institutions

It is, therefore, reasonable for preservation programmes to look for multiple strategies, especially if they are responsible for a range of materials over extended periods.

Conclusion

+ **The long-term strategies that are being practiced for digital preservation includes:**

  + **Use of standards for data encoding,**
  + **Structuring and description,**
  + **Emulation of obsolete software or hardware**
  + **Migration of data from one operating technology to another.**

+ **None of them have proven themselves against unknown threats over centuries of change.**

+ **Most of these strategies are, however, being used in the management of data, and it is likely that combinations of the will continue to be researched and proposed for large-scale, long-term preservation.**