Digital Archiving Strategies for the Long Term

e-Archiving for Posterity,
Antwerp, June 26 2003

Jean-François Blanchette
InterPARES 2 project
University of British Columbia

Archiving, digital or otherwise

- Three functions of archives:
  - Preservation of cultural heritage
  - Preservation of (documentary) evidence
  - To interpret/communicate archives for the present
- Preservation of physical carrier — e.g., temperature, relative humidity
- Preservation of ability to interpret linguistic encoding of documents
- Preservation of ability to interpret contextual dimensions of documents — e.g., diplomatics

The problem

- The transmission of digital information objects across technological boundaries (computer platforms, operating systems, applications) created by technological obsolescence
- A digital object possesses:
  – A physical dimension, as an inscription on a physical carrier (punch card, mag. tape, optical disc)
  – A logical dimension, as this inscription must be recognized and processed by software
  – A conceptual dimension, as an object produced and to be understood within a specific context

Physical Preservation

- Reliable method for maintaining data integrity in storage, including the need for
  - updates in storage systems
  - delivering data from storage to client
  - media refreshment/migration
- Given advances in storage technologies, may improve preservation
- Given reduction in costs of storage, may be more cost effective

Logical Preservation

- Determines how the inscription on a physical carrier is recognized by some application software, transformed into the system's memory and presented as an output
- The logical grammar of the inscription is independent of its physical realization on a carrier
- Grammar is based on data types, i.e., set of rules for representing digital information, primitive or composite
- Logical string, conforming to a data type, may be stored in a single or in multiple physical objects
- To preserve a logical object, we must know the requirements for correct processing of each object's data type and what software can perform it

Postscript

```
%!PS-Adobe-3.0
%%Title: (aae.doc)
%%CreationDate: (11:43 Lundi 16 octobre 2000 )
%%Pages: 1
%%Orientation: Portrait
%%EndComments
[...]
%%BeginFeature: *PageSize A4Small
     «/PageSize [595 842] /ImagingBBox null>> setpagedevice
%%EndFeature
%%EndSetup
%%Page: 1 1
[...]
gS 0 0 538 781 rC
86 75 :M
f57 sf
-.174(Longue vie \210 l\325acte)A
158 123 :M
-.192(authentique)A
143 171 :M
-.181(\216lectronique!)A
endp
showpage
%%Trailer
end
%%EOF
```

HTML

```
<HTML>
 <HEAD>
  <META CONTENT="text/html; charset=iso 8859-1">
  <META NAME="Generator" CONTENT="Microsoft Word 98">
 </HEAD>

 <BODY>
  <FONT FACE="Times" SIZE=7>
  <P ALIGN="CENTER">
  <A
HREF="http://www.internet.gouv.fr/pubs/acteauthentique.htm  l
">
  Longue vie &agrave; l&#146;acte authentique
&eacute;lectronique!
  </A>
  </P>
  </FONT>
 </BODY>
</HTML>
```

Conceptual preservation

- The object as we deal with it in the real world, an entity we would recognize as meaningful information produced within a specific context
- The same conceptual object may be represented by different logical encodings expressing different aspects of the same conceptual object, e.g., information processing (XML) vs. look-and-feel (TIFF)
- Different logical encodings of the same conceptual object can preserve its "essential characteristics" (TIFF, PDF)

Thus …

- In order to preserve a digital object, we must be able to identify and retrieve all of its digital components, i.e., the logical and physical objects necessary to **reconstitute** the conceptual object
- That is, to access any digital object, **stored** bit sequences must be **interpreted** as logical objects and **presented** as conceptual objects
- In the paper-and-ink world, the basis of preservation is the caring for the integrity of the physical carrier itself, but…

… counter-intuitively…

- Digital preservation is not a simple process of preserving physical objects (stored bit sequences), but one of **preserving the ability to reproduce the objects,** and this process is complete only when the objects are successfully output!
- Preserving a digital object **does not imply** preserving its physical and logical components and their relationships without alteration!
- Archives are **not** simply "a neutral communication channel for transmitting information to the future, which does not corrupt or change the messages transmitted in any way."

Reframing the problem

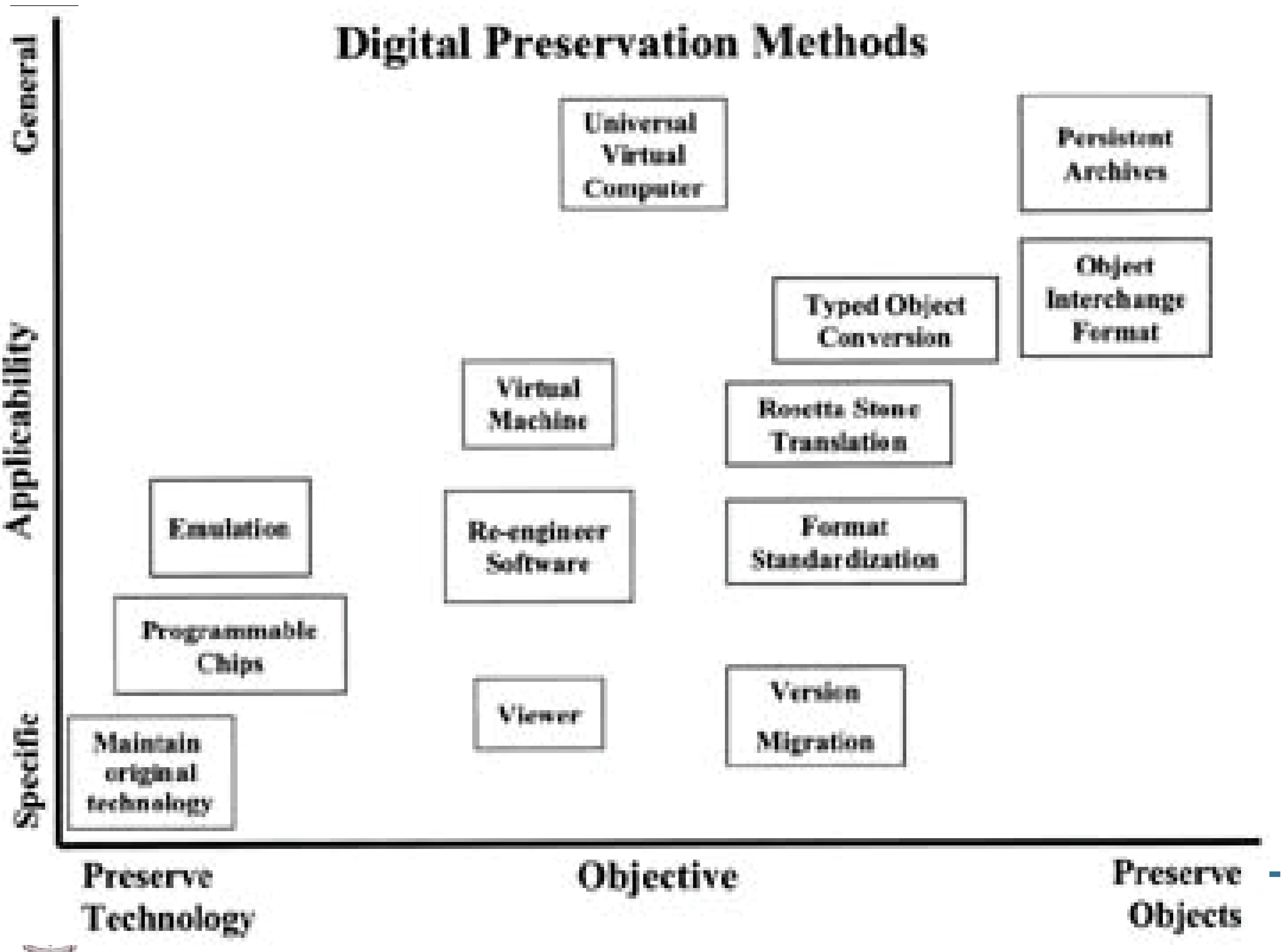
- The problem becomes, "Which changes are permissible and/or beneficial?"
- Given that a digital information object is something that can only be **re-constructed** by using software to process stored inscriptions, it is necessary to have an **explicit model or standard** that provides a **criteria for assessing the authenticity** of the re-constructed object
- InterPARES 1 has produced such criteria for electronic records:
  - **Benchmark Requirements** for creation, maintenance, and handling of active records
  - **Baseline Requirements** for copies of inactive records

Which technology?

- Any technological solution to digital information preservation should satisfy the following criteria:
- **Feasibility**: hardware/software must exist for the method
- **Sustainability**: must be applicable in the future
- **Practicality**: reasonable difficulty and expense
- **Appropriateness**: preservation needs must be determined on the basis of a specific definition of the essential characteristics of the object to be preserved. E.g., in the case of web sites, should we preserve:
  - The "behavior" of the site, i.e., hyperlinks, applets, etc.
  - Or individual pages?

The spectrum of preservation

- **(A) Preserve technology:** keep data in original logical/physical formats and use technology associated with those formats (media drivers, viewers) to access the data and reproduce the formats
- **(B) Update as-you-go:** migrate data formats as technology changes, thus using state-of-the-art technology for storage/access/output
- **(C) Preserve conceptual objects:** focus on preserving the essential characteristics of objects, defined explicitly and independently of specific hardware/software


Digital Preservation Methods
General
Applicability
Specific
Universal Virtual Computer
Persistent Archives
Object Interchange Format
Typed Object Conversion
Virtual Machine
Rosetta Stone Translation
Emulation
Re-engineer Software
Format Standardization
Programmable Chips
Viewer
Version Migration
Maintain original technology
Preserve Technology
Objective
Preserve Objects

(A) Preserve technology

- Create IT museums, keeping media drivers, hardware and software platforms running for as long as we need to read data …
- **Pros:**
  – Archival theory doesn't have to rethink itself
- **Cons:**
  – Fails the sustainability and the practicality tests

(A) Emulation *à la* Rothenberg

- Each computing platform is *emulated* by the succeeding generation of computing technologies – F(E(D(C(B(A( ))))))
- Proven concept: Virtual PC for Macintosh, Videogames
- **Pros:**
  - Preserves look-and-feel of computing environment
  - Preserves functionalities of software
- **Cons:**
  - Impossibly complex on a comprehensive scale
  - Implies providing user-support for preceding generations of software

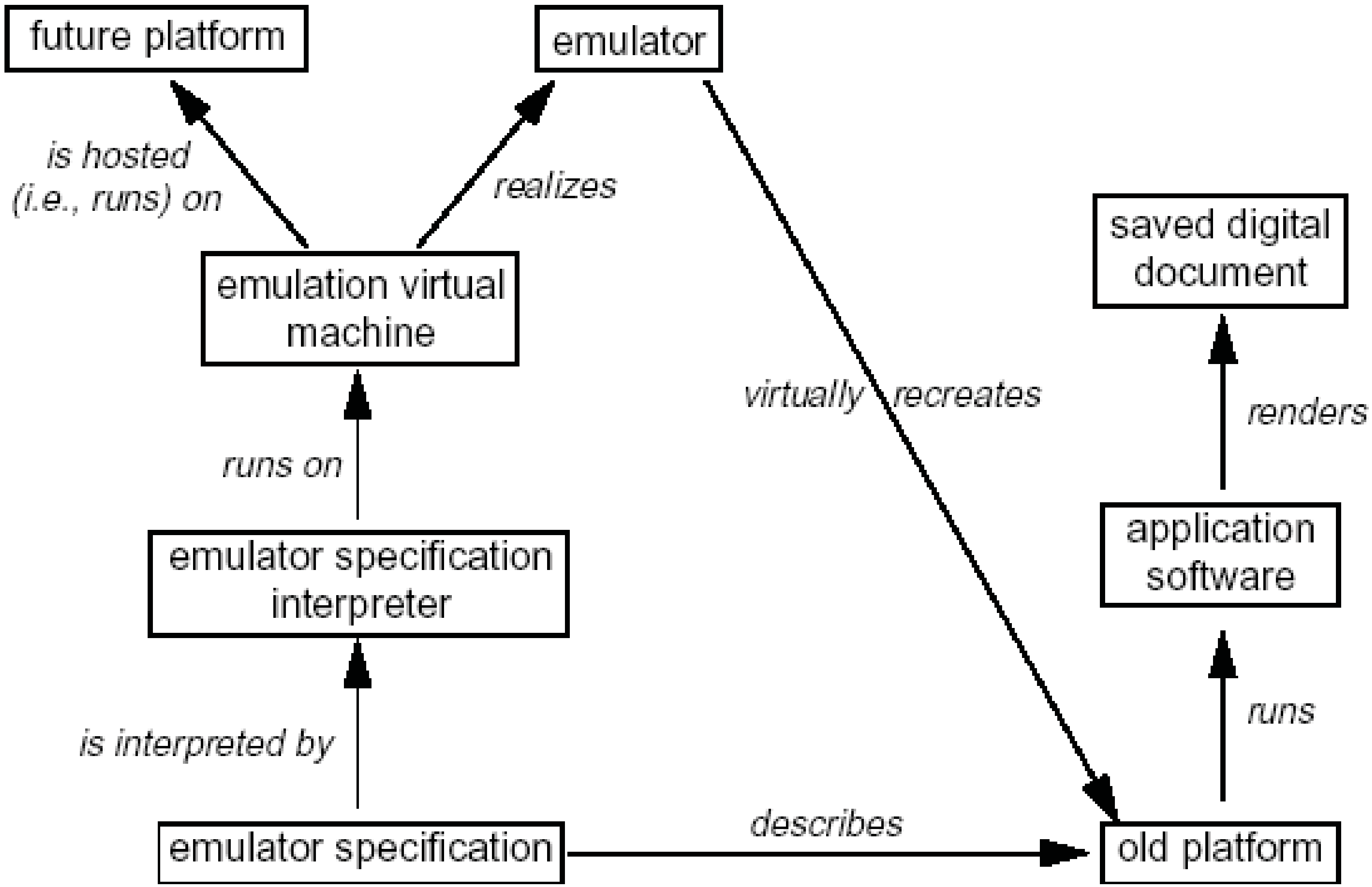
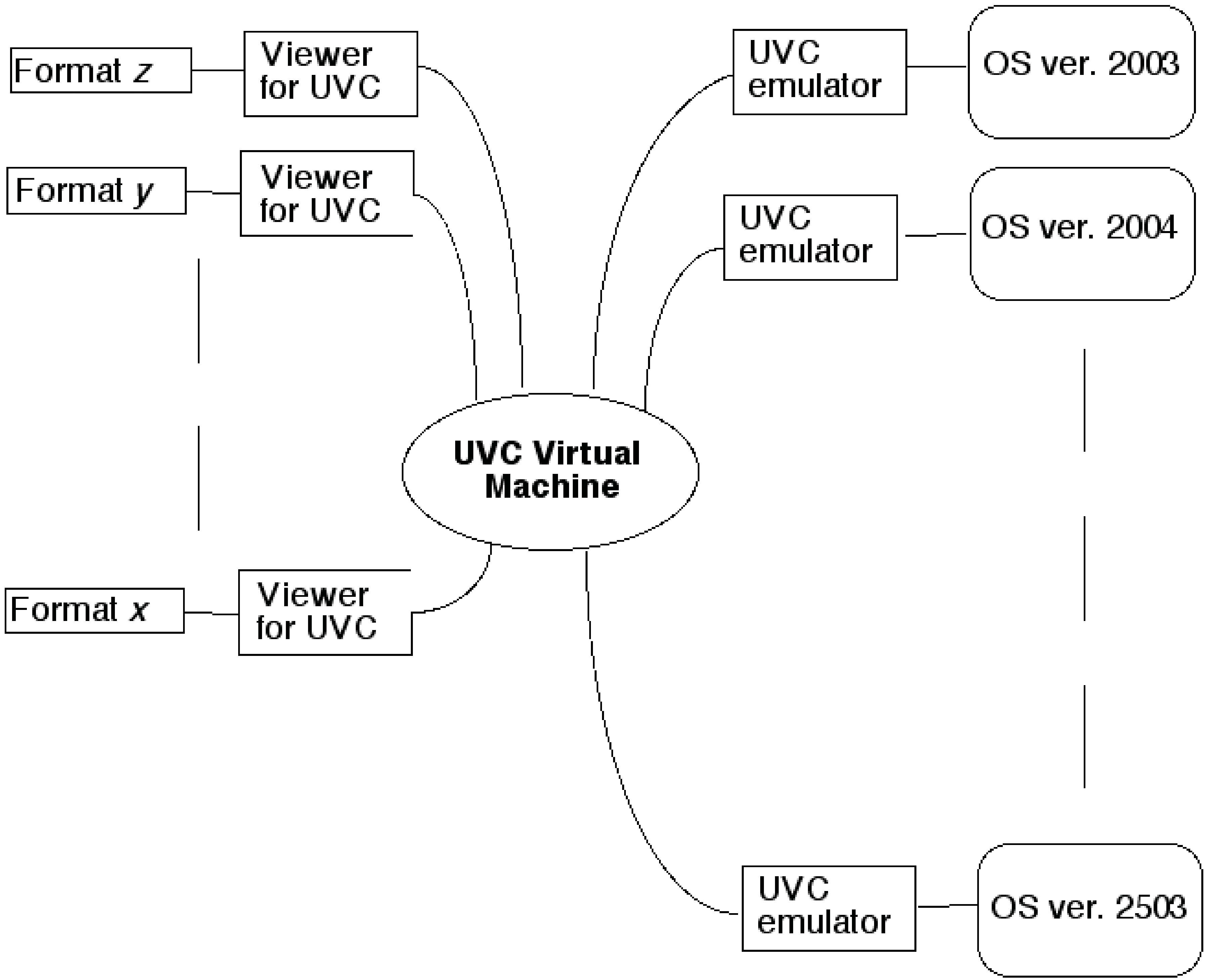


Figure 3: Elements of emulation-based preservation

(A) Emulation *à la* Lorie

- Specifies a *Universal Virtual Machine* (UVC), capable of performing essential algorithmic functions
- Digital objects are preserved in their original formats, along with encoding/decoding rules for the UVC
- Computer system vendors commit to creating a UVC emulator on all future platforms
- The engineering burdens of emulation are distributed among actors, but performance is likely to be poor



Format *z*
Viewer for UVC
Format *y*
Viewer for UVC
Format *x*
Viewer for UVC
UVC Virtual Machine
UVC emulator
OS ver. 2003
UVC emulator
OS ver. 2004
UVC emulator
OS ver. 2503

(A) VERS

- Victorian Electronic Records System, concentrates solely on documents and the preservation of their look-and-feel:
- (1) Migrate everything to PDF
- (2) Preserve
- (3) Trust that current PDF specifications are complete
- (4) Trust that a PDF viewer can be engineered from specifications for every future computing platform to come

(B) Version Migration

- Within the same family of products or data types, software vendors supply conversion routines so that newer versions of products can read older formats
- **Pros:**
  - Do-it-yourself digital preservation: we are all familiar with it …
- **Cons:**
  - No explicit user control of the process
  - Endows older formats with attributes they might have never possessed in the first place

(B) Format standardization

- Transform various data types to single (supra) standard type:
  - plain text for all textual documents
  - bitmaps for all visual documents
  - tab-delimited arrays for databases, etc.
- **Pros:**
  - Using the lowest-common denominator gives better assurance of ability to process data in the future
- **Cons:**
  - But even standards evolve, e.g., EBCDIC to ASCII to Unicode

(B) Rosetta Stone Conversion

- (1) Create a sample set of data objects which cover all characteristics of the source format
- (2) Create a reference set of what objects in sample should output like, e.g., on microfilm or paper
- (3) Given the reference set, create a target sample set in target format
- (4) Comparing target sample with original sample, deduce the rules for translation
- **Pros:**
  - Translations are always performed from original format, avoiding all intermediate migrations
- **Cons:**
  - Unlikely to work on complex digital objects

(C) Object Interchange Format

- Define information objects at the conceptual level, formally specify them and articulate corresponding logical model (e.g., DTD in XML)
- The models serve as bridges between heterogeneous systems and data types, enabling greater exchange
- To preserve, build interpreters enabling target systems in the future to import objects in such formats
- **Pros:**
  - Essential properties of objects defined by experts with substantial knowledge of their creation and use, thus embedding domain knowledge in their transmission across space, time, and technologies
- **Cons:**
  - Is not designed for preservation as such, i.e., XML also likely to evolve and develop proprietary extensions

(C) Persistent Archives

- Comprehensive framework integrating very large DB technology, digital libraries for access, and archival concepts for preservation – See Prof. Underwood's presentation later today

Conclusion

- Any solution must be **evolutionary**:
  - continuing changes in the nature of the problem
  - continuing escalation of user demands
- If the preservation solution cannot grow and adapt, **the solution itself will become obsolete**
- Technological frameworks which ensure integrity solely through cryptographic technologies (digital signatures or hash functions) must confront the archival perspective
- Legislative and regulatory frameworks which have reformed evidence law around such specific technologies are **not** technologicaly neutral and may face rapid obsolescence

References

- Kenneth Thibodeau's article: "Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years", http://www.clir.org
- Very good reference site on digital preservation: http://www.nla.gov.au/padi/
- InterPARES: http://www.interpares.org
- Questions/comments:

  Jean-Francois.Blanchette@ubc.ca