Digital Preservation Summary

Jeff Rothenberg

Jeff_Rothenberg@acm.org (310/664-1967)

April 4, 2003

Digital records are very vulnerable to loss

- **Media decay or "evaporation" of bits**
  - **Due to physical, chemical, magnetic effects, etc.**
- **Media obsolescence**
  - **Physical and logical format incompatibilities**
  - **Unavailability of suitable "drives" or "controllers"**
- **Dependence on incompatible or obsolete software**
  - **e.g., for word processing or hypermedia documents, DBs, etc.**
- **Dependence on obsolete software environments**
  - **Unavailability of OS, I/O drivers, etc. for required software**
- **Dependence on obsolete hardware**
  - **Unavailability of hardware required to run required software**

So how long will digital records last?

- **Forever?**
  - **Because they can be copied perfectly (i.e., proliferate without degrading)?**
- **No!**
  - **Because of the vulnerabilities discussed above, the best we can say is...**
- **“Digital records last forever — or five years, whichever comes first”**

Solving the media problem is “straightforward”

- **Truly “archival” digital storage media are not yet cost-effective**
  - **Since media (and their formats & reading devices) become obsolete so fast**
  - **And storage capacity, density, & speed increase with each new generation**
  - **The market will not pay for long-lived media while this progression continues**

- **So, must copy records to new media while still readable**
  - **The same as for non-digital records**
  - **However, must take into account obsolescence as well as physical lifetime**

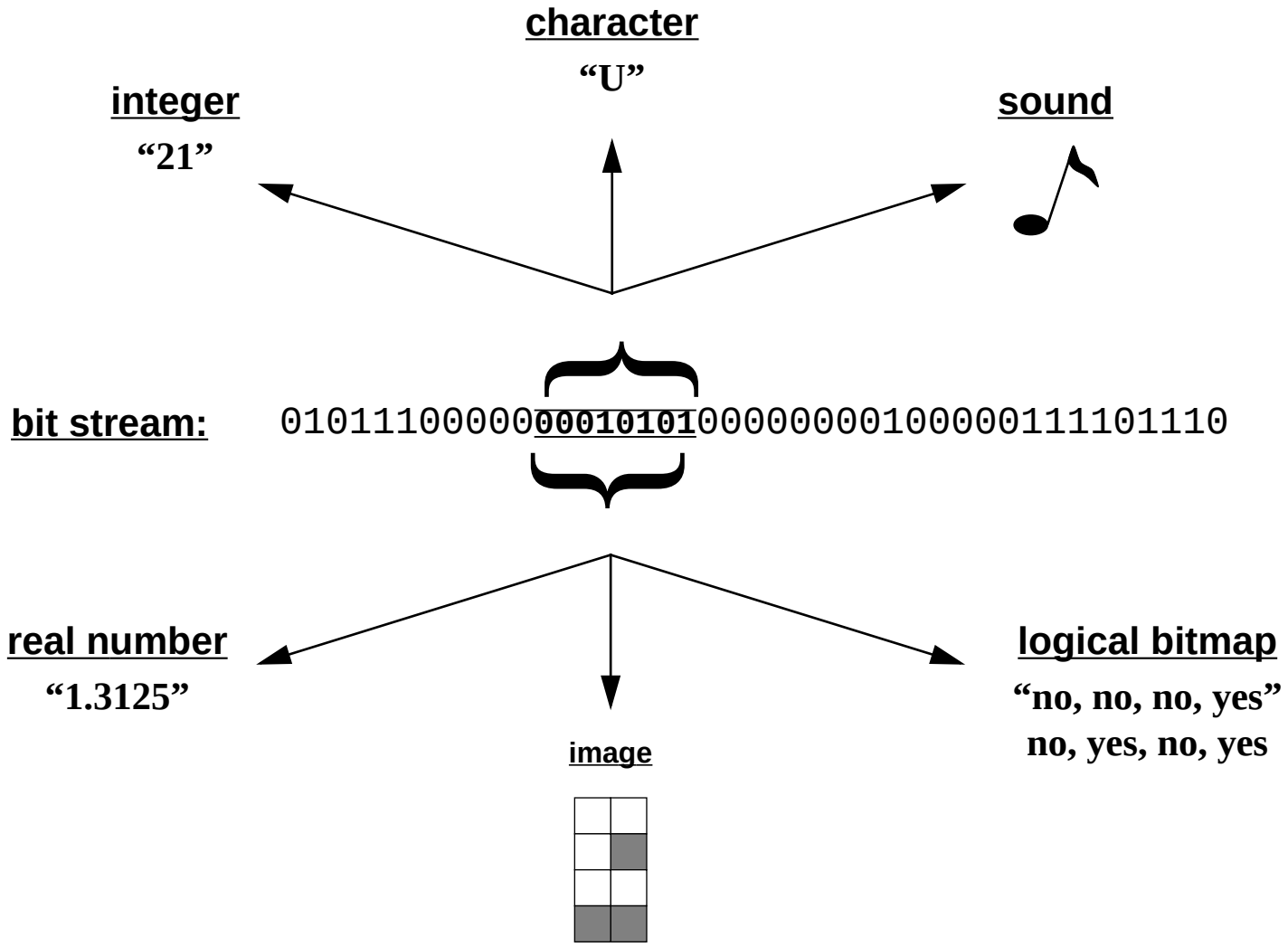
But all digital records are software-dependent

- **Digital records can be seen only by running a program**
  - **They are stored in encoded form, understood only by a program**
  - **They cannot be accessed, read, or printed without that program**
  - **They must be interpreted to be made intelligible to a human**
  - **They are essentially programs**
  - **Examples: ASCII character stream, hypermedia, database, animated film, interactive video game**

- **The data file for a software-dependent record is *not* enough**
  - **The file can be properly interpreted only by its software**
  - **Without the software, the record is unusable (may not even really exist)**
  - **"Virtual records" may consist of multiple (distributed) files**

- **Software-dependent records are really *system*-dependent**
  - **They require a software environment (OS, drivers, etc.)**
  - **Which in turn requires a hardware environment (CPU, I/O devices, etc.)**

Bits in a bit stream can represent *anything*



Furthermore, many new records are *inherently digital*

- ***Inherently digital*** **records are those whose meaning or usability arise from and rely on their being encoded in digital form**
- **They cannot be meaningfully represented as page images**
  - **Doing so loses essential aspects of their contents and/or behavior**
- **Examples include dynamic, active or interactive artifacts**
  - **Multimedia (e.g., web pages, CD-ROM publications, Ph.D. dissertations)**
  - **Generated dynamically (e.g., calendars, agendas, bookkeeping data)**
  - **Generated on request (e.g., customized weather maps)**
  - **Generated automatically (e.g., JavaScript, cgi, ASP web pages, servelets)**
  - **Active presentation (e.g., animation, simulation, virtual reality)**
  - **Databases (where transactions update relationships and inferences)**
  - **Interactive (e.g., applets, interactive virtual reality)**

A particular “view” of information may be crucial

Example: Space Shuttle O-ring damage vs. temperature Prior to the Challenger disaster

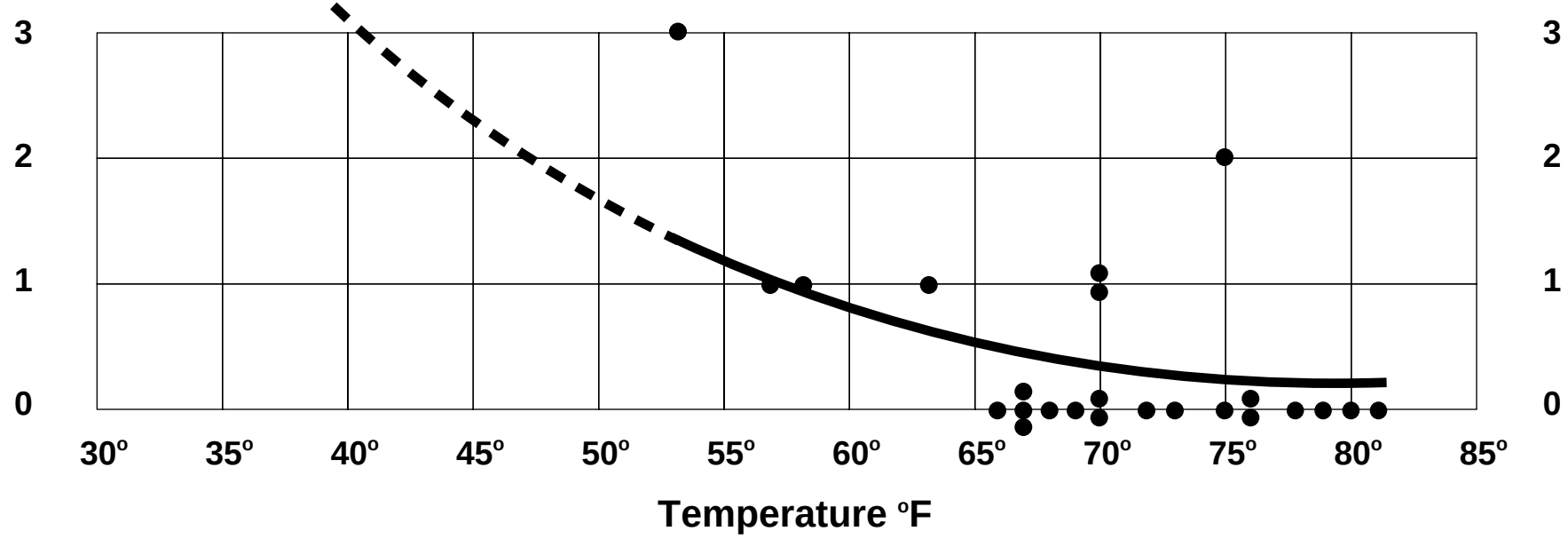
Levels of O-ring damage																	
3	1																
2												1					
1		1	1	1					2								
0					1	3	1	1	2	1	1	1	2	1	1	1	1
	53	57	58	63	66	67	68	69	70	72	73	75	76	78	79	80	81

Temperature °F

Revealing view of Space Shuttle O-ring Data

Extrapolation of damage curve to the 31° F temperature forecast for Challenger's launch on January 28, 1986.

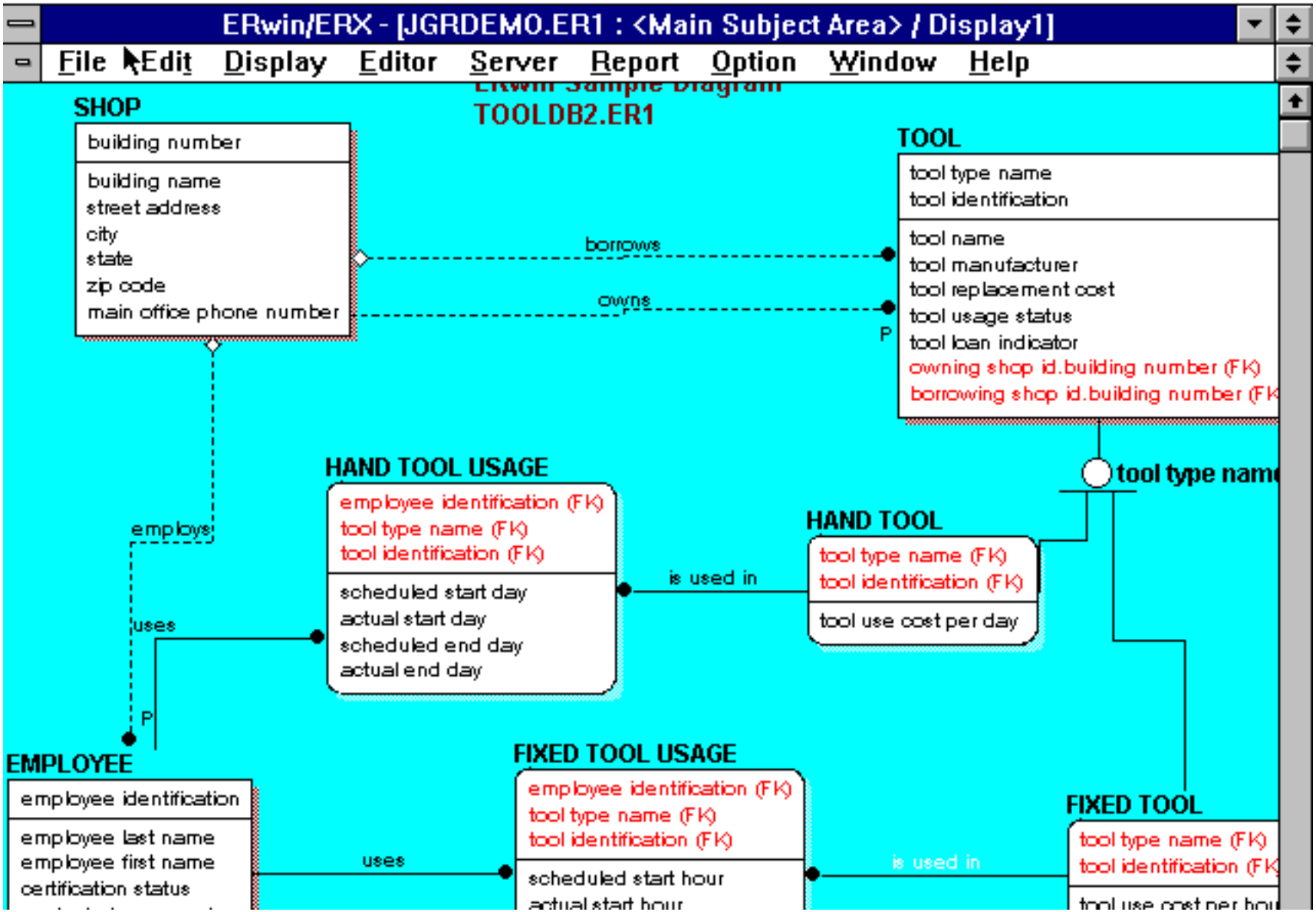
Dots indicate temperature and O-ring damage for 24 successful launches prior to Challenger. Curve shows that increasing damage is related to cooler temperature.



Every digital record is really a program

- A *program*
  - Is a sequence of commands in some formal language
  - That is intended to be interpreted
  - By an interpreter that understands that language
- An *interpreter*
  - Is an active process
  - That knows how to perform commands
  - Specified in a given formal language
- Interpretation ultimately involves hardware
  - ASCII codes are rendered by a printer or display
  - More complex entities are interpreted by software (applications)
  - But all S/W is ultimately interpreted by hardware

What you see may *not* be what you get



Text may not tell the story at all

```
V2.24 ERwin
 if
  %JoinPKPK(oldrows,newrows," <> "," or ")
 then
  select count(*) into numrows
   from %Child
   where
    %JoinFKPK(%Child,oldrows," = "," and");
  if (numrows > 0)
  then
   signal parent_updrstrct_err
  end if;
 end if;
 if
  %JoinPKPK(oldrows,newrows," <> "," or ")
 then
  update %Child
   set
    %JoinFKPK(%Child,newrows," = "","")
   where
    %JoinFKPK(%Child,oldrows," = "," and");
 end if;
```

Saving the bits is necessary but not sufficient

- **Saving the bit stream of a record without saving its interpreter**
  - **Is like saving hieroglyphics without saving a Rosetta Stone**
- **But *worse*, since an interpreter is not just another record**
  - **It is software**
  - **Which must be executed (i.e., interpreted)**
  - **And the record must still be understood (i.e., further "interpreted")**
- **So digital records are generally software-dependent**
  - **And software is ultimately hardware dependent**

So why is it hard to preserve digital records?

- **Can't "just save" digital records like physical records**
  - **The medium carries all attributes of a traditional record**
- **Digital records require an extra "interpretation" step**
  - **To be made human-readable**
  - **Especially if they are dynamic, responsive, interactive, or "active" (executable)**
  - **But even simple text formats require interpretation**
- **An interpreter can be hardware or software**
  - **But hardware is limited to interpreting simple, static languages**
  - **And must be well-specified in order to be built**
  - **Whereas software can interpret more complex, dynamic languages**
  - **And need not be well-specified to run**
- **So most digital records rely on software interpreters**
  - **E.g., application programs**
  - **Which become obsolete**
- **And executing software requires hardware**
  - **Which become obsolete**

Overview of proposed approaches to preservation

- **Non-solutions**
  - **Do nothing**
  - **Digital archaeology**

- **Partial solutions**
  - **Save page-images of artifacts**
  - **Extract and save “core contents” of artifacts**
  - **Translate artifacts into standard or “canonical” forms (without migration)**
  - **Rely on “viewer” programs to render obsolete formats in the future**
  - **Save metadata to help interpret saved bit streams (“assisted archaeology”)**
  - **Save source-code of rendering software (for future reverse-engineering)**

- **Potentially complete solutions**
  - **Formalization (replace artifacts by formal descriptions of themselves)**
  - **Migration (repeatedly convert artifacts into new formats)**
  - **Emulation (run original rendering software on virtually recreated hardware)**

Standards are not enough

- **Ultimate standards are not realistic in the foreseeable future**
  - **Information science is still inventing itself**
  - **Even the categories of kinds of information processing are not yet clear**
  - **So ultimate standardization is premature**
- **Using successive, evolving standards would require translation**
  - **But translation between standards is rarely reversible without loss**
  - **So this cannot reconstruct an original artifact**
  - **Translation forward across "paradigm shifts" may be impossible**
  - **So old artifacts may eventually be abandoned or corrupted**
- **Evolving standards will always lag behind state-of-the-art use**
  - **Until information science stops evolving**
  - **So state-of-the-art artifacts are likely to be "orphaned"**
- **Can't force users to conform to constraining standards**
  - **This asks them to forego the use of new capabilities**
  - **Which are the motivation for using information technology in the first place**

Formalization is very difficult

- **A formal description of a saved digital artifact's logical format**
  - **Would allow properly interpreting that format in the future**
  - **So long as the formal description itself remained understandable**
  - **This would allow properly rendering the saved digital artifact**
  - **Without running its original software**

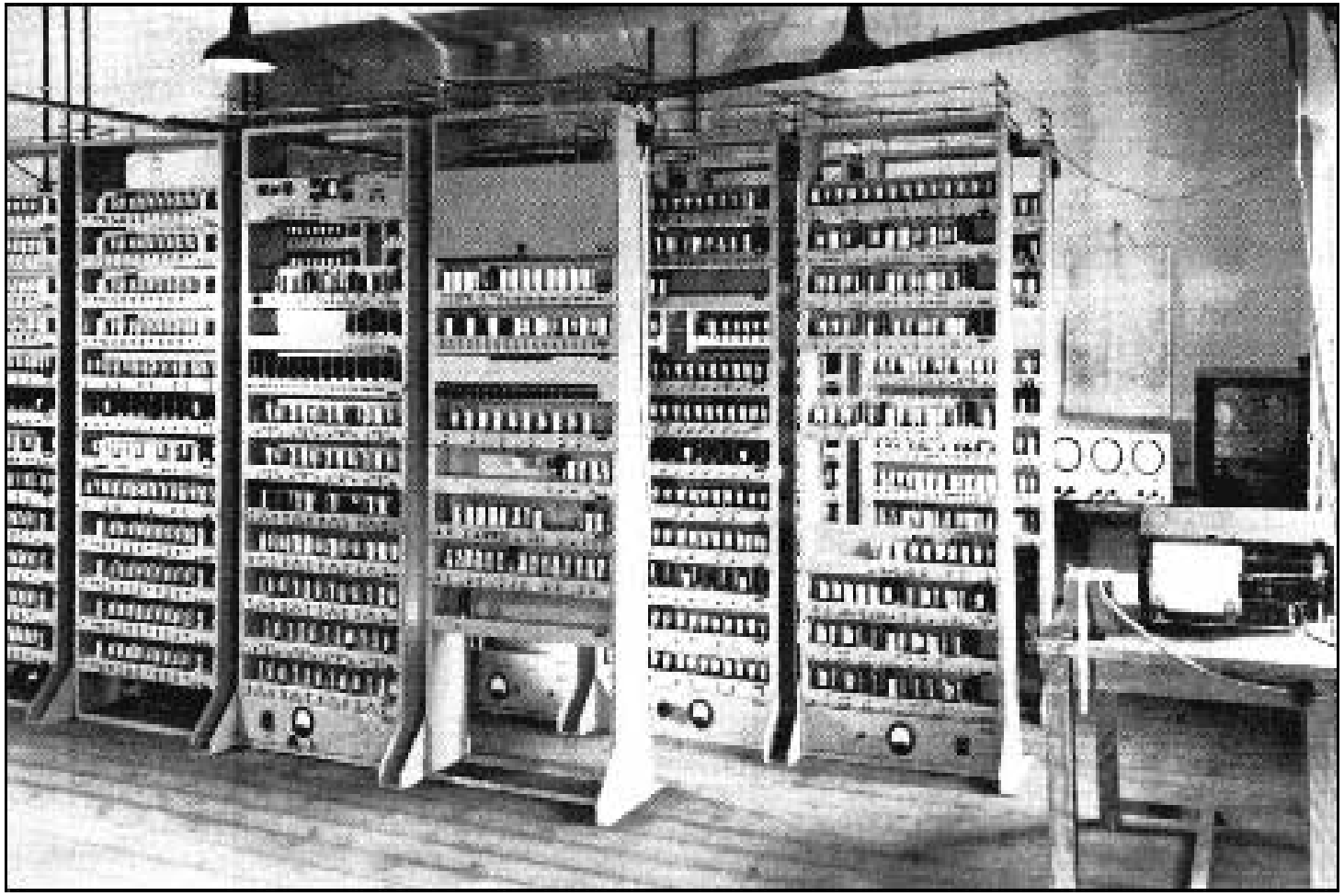
- **Unfortunately computer science cannot do this very well yet**
  - **Even for well-documented, well-defined formats**
  - **Let alone for arbitrary, new, proprietary formats**

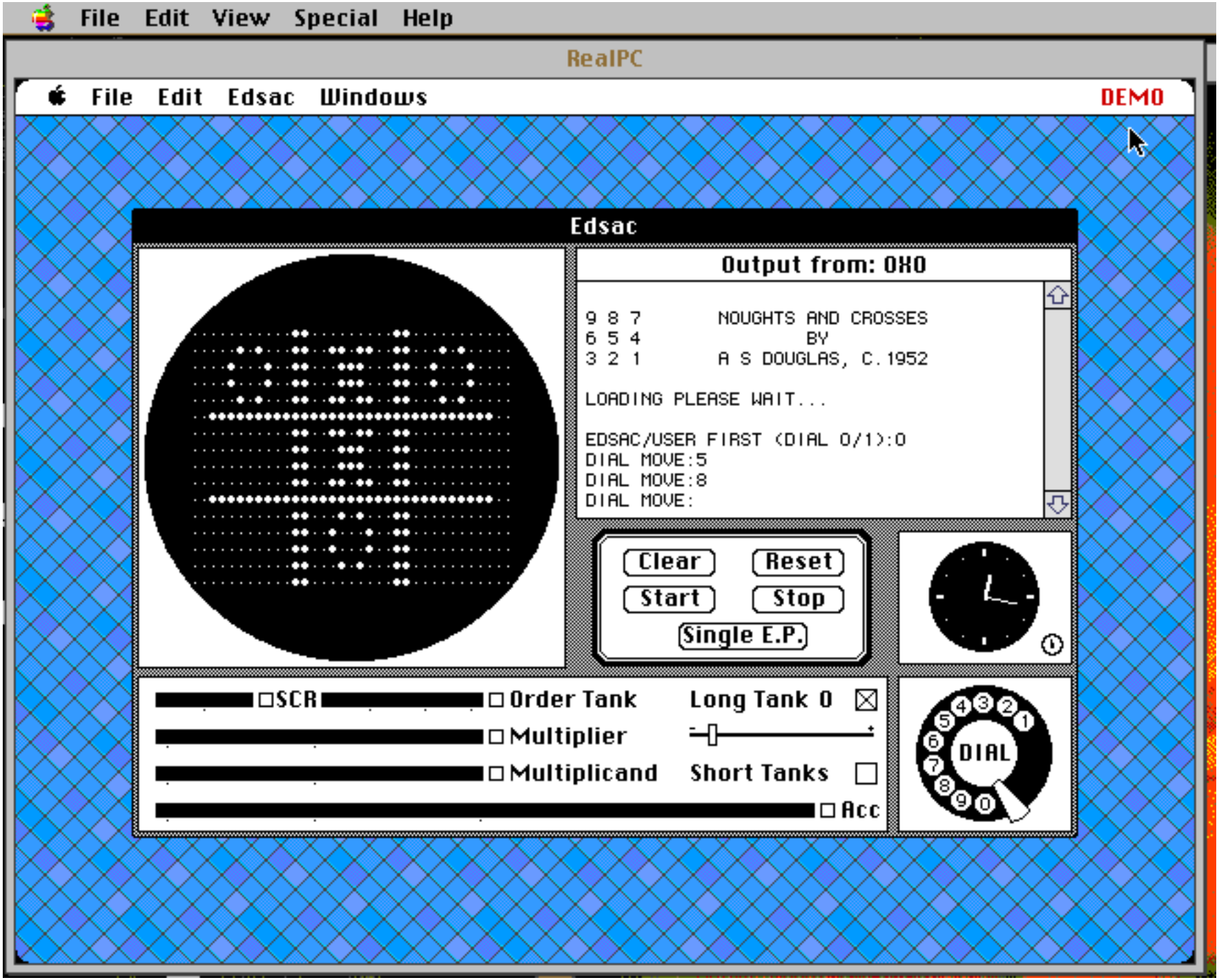
- **The only complete description of a format is its interpreter**
  - **i.e., the software that knows how to render it**

Emulation is the only proposed approach that...

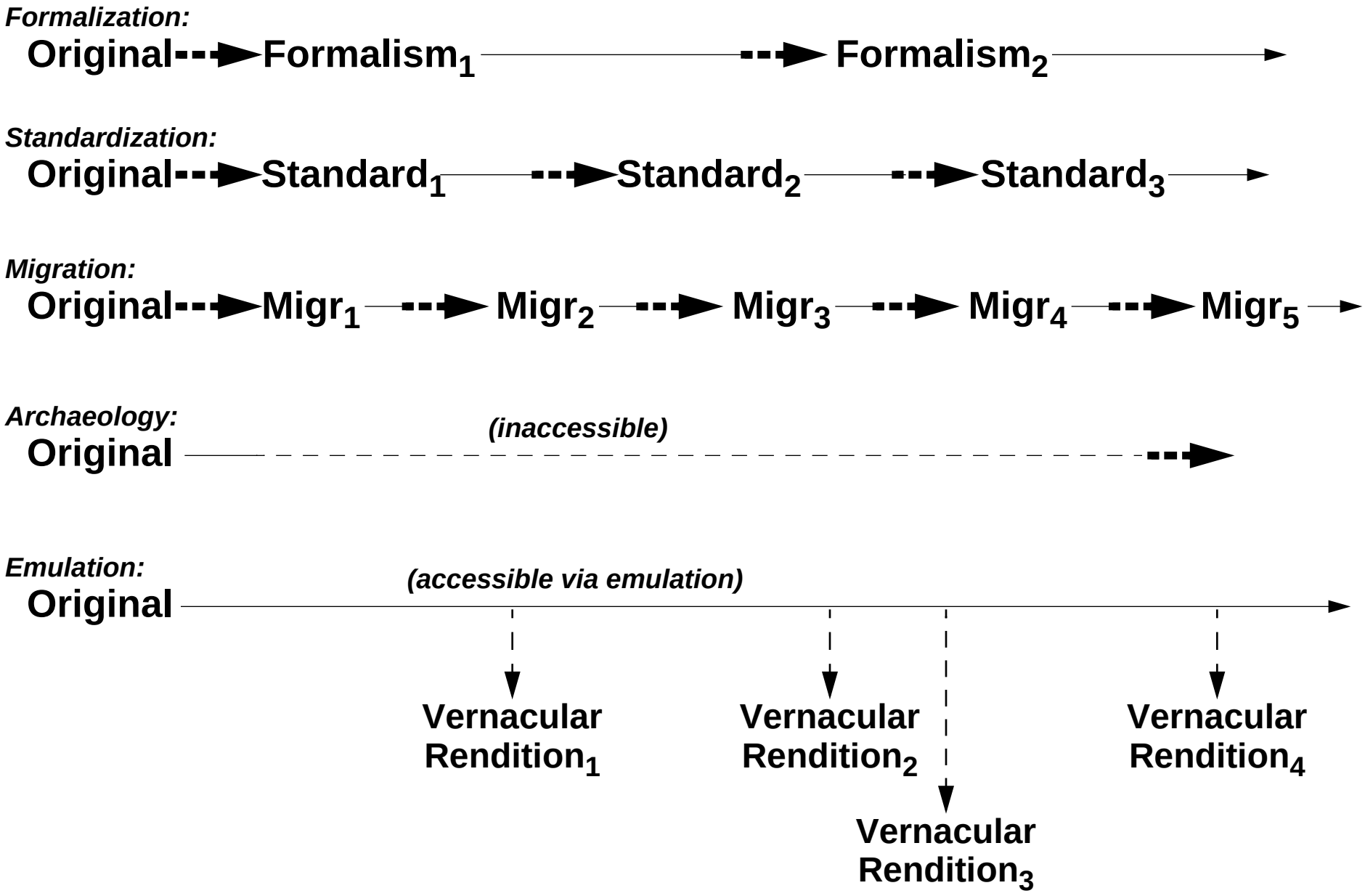
- **Can potentially preserve “digital-originals”**
- **Can preserve executable digital artifacts (i.e., “behavioral preservation”)**
- **Can preserve all kinds of digital artifacts in a single, consistent way**
- **Obviates the need to understand the formats of individual records**
  - **Except what software is needed to view them**
- **Requires zero per-record (artifact) effort, both initially and over time**
  - **Except for copying bitstreams onto new storage media**
- **Defers the need to convert records into new formats unless and until it is desired to access them in such formats in the future**

EDSAC: the first electronic digital computer




File Edit View Special Help
RealPC
File Edit Edsac Windows
DEMO
Edsac
Output from: 0X0
9 8 7
6 5 4
3 2 1
NOUGHTS AND CROSSES
BY
A S DOUGLAS, C.1952
LOADING PLEASE WAIT...
EDSAC/USER FIRST (DIAL 0/1):0
DIAL MOVE:5
DIAL MOVE:8
DIAL MOVE:
Clear
Reset
Start
Stop
Single E.P.
SCR
Order Tank
Long Tank 0
Multiplier
Multiplicand
Short Tanks
Acc
DIAL

Process models of preservation approaches



Mixed preservation strategies

- **Traditional conservation is both medium-specific & discipline-specific**
  - **Treat books differently from paintings, sculpture, textiles, furniture, audio tape, etc.**
  - **Libraries, archives, museums, scientific repositories, etc. have different agendas**
- **But the homogeneity of all digital artifacts creates new possibilities**
  - **All digital artifacts can be treated by any of the approaches we have discussed**
  - ***And* these approaches are not mutually exclusive**
- **Using one approach for everything would be simpler**
  - **May be unwarranted or too expensive for some kinds of records**
  - ***But* it would take advantage of economy of scale, so might ultimately be cheaper**
  - **Using one cheap approach would be better than using many expensive ones!**
- **For now, consider using multiple approaches in parallel:**
  - **Digital archaeology (for records that are unlikely to be accessed)**
  - **Page-image techniques (for simple records)**
  - **Formal methods (when applicable and affordable)**
  - **Standards (when available)**
  - **Migration (when it needs to be done anyway, i.e., for active records; *or* as a stopgap)**
  - **Emulation (if original behavior is needed; *or* as a cheap backup, to preserve everything)**