Principles for
100-year Digital Preservation

H.M. Gladney

Most content starts in digital form.
Some content is fundamentally digital.
People want to save some of it for the long term.

What's needed to preserve bit-streams?
What will be needed to interpret saved bit-streams?
What can be done to make the content trustworthy?
What’s the Problem?

Digital infrastructure in its infancy; paper infrastructure has been slowly refined for over 3000 years
Society has had little commitment to digital preservation
Neither U.S. NARA nor the Library of Congress fills the gap
LoC NDIIPP is narrowly focused, and has made little technical progress
NARA doesn’t archive broadly, just federal government records

“The state of affairs in 1998 could easily be summarized: *

- proven methods for preserving … electronic records were limited to the simplest forms of digital objects;
- proven methods [could not be] scaled … [for] expected growth; and
- archival science had not responded to the challenge of electronic records sufficiently to provide a sound intellectual foundation for articulating archival policies, strategies, and standards for electronic records.”

* Thibodeau (NARA) 2002, Overview of Technological Approaches to Digital Preservation …
Scope Limitations and Status

Digital Preservation $\equiv$ mitigation of the deleterious effects of technology obsolescence, media degradation, and fading human memory.

Technical components only (e.g., not addressing professional education needed).

Increment to available document and digital repository technology.

Focus: preserving a small set of closely related info.

- Caveat: no set of information is truly bounded!

Unsolved challenge: almost nobody provides adequate metadata with their works—not even for works that cost large efforts and expense—after a decade of metadata research and design.

Every conceptual challenge answered; no criticism received.

Software design, prototypes, and pilot is the next step.
Emphasis in TDO† Approach

Divide and conquer

Assume the worst and try to mitigate its potential effects

- Repositories are lost. Institutions disappear with their content.
- Clever rogues try to alter valuable information feloniously.
- Disgruntled, careless, or stupid employees defeat good procedures.
- The last guy who knows critical information just died.

Focus on information objects. End users don’t care how archives work, just that they deliver what was stored.

Provide for every kind of govt., business, home, and educational situation, and for every significant file type.

With a “worst case” solution in hand, consider whether less expensive means would work for simpler cases.

† TDO = Trustworthy Digital Object
Preservation is a Different Topic than Repository Management

NARA’s challenge is atypical

- Huge data volume from a huge bureaucracy
- Intellectual property challenges are secondary
- Business records tended by agency records administrators
- Large record sets sharing format and metadata

Scholarly works and scientific records in academia

- No-one stands between archivists and information providers
- “Designated communities”, e.g., university students and faculty
- Each proposed accession might need individual handling
- Funding for preservation an unresolved challenge

⇒ Different challenges today and different workflows
Workflow: Bureaucratic vs. Cultural
### From Workshop Prospectus: Digital Archiving Models, Representation Languages and Standards

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canonical representation for archiving?</td>
<td>TDO (Trustworthy Digital Object) structure</td>
</tr>
<tr>
<td>How to compress data?</td>
<td>Not a digital preservation topic.</td>
</tr>
<tr>
<td>How to manage interoperability?</td>
<td>Current EDP standards.</td>
</tr>
<tr>
<td>How to prepare content for archival submission.</td>
<td>(1) Add metadata that archivists want.</td>
</tr>
<tr>
<td></td>
<td>(2) Encode bit-strings to be durable.</td>
</tr>
<tr>
<td></td>
<td>(3) Sign and seal content packages.</td>
</tr>
<tr>
<td>Trustworthiness of archived information?</td>
<td>Durable Evidence (this talk)</td>
</tr>
<tr>
<td>Role of standards in information packages? OAIS packages and submission</td>
<td>Starting point for Durable Encoding.</td>
</tr>
<tr>
<td></td>
<td>TDO format (this talk)</td>
</tr>
<tr>
<td>Taxonomies, Thesauri and Ontologies</td>
<td>Not a preservation topic.</td>
</tr>
<tr>
<td>Achieving Semantics Interoperability</td>
<td>“the Holy Grail of knowledge management”</td>
</tr>
</tbody>
</table>

© 2006, H.M. Gladney, HMG Consulting  
http://home.pacbell.net/hgladney/
## From Workshop Prospectus:

### Issues in Manufacturing Engineering Informatics

<table>
<thead>
<tr>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is to be archived?</td>
<td>Almost every widely used file format will occur among manufacturing information—even file formats for music and musical performances</td>
</tr>
<tr>
<td>How is this information to be represented?</td>
<td>Of course, but there are surely others of comparable significance</td>
</tr>
<tr>
<td>Is STEP a starting point for content information?</td>
<td>Content management topic, not preservation</td>
</tr>
<tr>
<td>How to scale from parts to systems?</td>
<td>Existing EDP standards, perhaps from IEEE?</td>
</tr>
<tr>
<td>How to incorporate tolerance information?</td>
<td>Look to recent task groups such as PREMIS</td>
</tr>
<tr>
<td>Initial draft for Preservation Description Information?</td>
<td></td>
</tr>
<tr>
<td>How to retrieve product data?</td>
<td>Content management topics, not preservation</td>
</tr>
<tr>
<td>Generic features and contextual indexing?</td>
<td></td>
</tr>
</tbody>
</table>
Preservation: a special case of Communication

Information consumer cannot obtain producer’s explanations.

Talk about the users, not about repositories!
Distinguishing current needs is politically pragmatic.
An Information Communication Model

Can describe transformations objectively, except for 0→1 and 9→10. What the information consumer receives should not depend on the transmission path from the information producer. (Implies an OAIS DIP to OAIS SIP relationship.)
Which object are we talking about?

5 cannot be identical to 4, but 3' can be created identical to 3.

§ Thibodeau, *Overview of Technological Approaches to Digital Preservation* ..., CLIR, 2002.

© 2006, H.M. Gladney, HMG Consulting http://home.pacbell.net/hgladney/
Prominent Epistemological Issues

“Trusted Digital Repositories” are problematical†
(misleading and unachievable)

People are confused about “dynamic digital objects” *
Accidental info. tends to obscure what is essential.
There are limits to what can be communicated.
Carefully distinguish subjective factors from objective.
Knowledge is much more than information.

Every object identifies a collection; every collection
identifies objects ➔ one structural schema is enough!

Ternary relationships and references are key constructs.

* Duranti, Documents Numérique 8(1), 1-14, 2004
§ Thibodeau, Overview of Technological Approaches to Digital Preservation …, CLIR, 2002.
Part of the Context: Nested Digital Repositories

“Repository” is ambiguous!


© 2006, H.M. Gladney, HMG Consulting  http://home.pacbell.net/hgladney/
Information Object Classes

Don’t forget surrogates for material objects! (e.g., automobiles)
Digital Object (or a Collection of Anything)

Tightly bound package.
Including metadata.
Linked schema & standards.
Nested versions.
Using ternary relationships.
Saving every important version.
Recursive and grounded.
Names (UUIDs) and Relationships

What kind of preserved object identifiers do we want?
- Universally unique forever.
- Human-readable is not a requirement.
- Not dependent on naming authorities.
- Without collisions with legacy naming schemes.
- To identify collections, documents, and material objects.
- To objects, into objects (as bookmarks), or to object extents.

Ternary relationships: the quintessential structuring tool!
- Can say anything that can be said objectively!†
- Are the elements of structure graphs.
- Easier to manipulate (e.g., with SQL) than binary relationships.

Bit-string Durability: Replication (as for books)

LOCKSS* successful, but embeds too much policy
Avatar implementing rules for replication

► Generalize
Design rules for replication
Design rules for access control

Durable Evidence for Info. Packages

UUIDs for digital objects, for material objects, and for object collections
Relationships whose elements are identifiers and scalars
Certificates with recursion grounded in a few institutions
Share keys by personal connections
All packaged with XML
Signed and sealed just as in the age-old practice (Japanese ukiyo-e)

Durable Encoding for Content Bitstrings

Starting point is a few widely understood standards
- Unicode/UTF, small XML core, BNF, ASN.1, identifier syntax

Universal Virtual Machine (enhanced Turing Machine)
- Bit-addressed, any number of registers, condition codes
- Segmented memory—some segments private, others shared
- Input/output streams,
- Still needs multiprogramming and real-time instructions

Need one UVC emulator for each HW architecture

Need one UVC application for each file type (~1000 types)

Don’t save anything irrelevant, e.g., O/S portions

Save the original file together with UVC application

Durable Encoding for Data Content Blobs

Objective: **render** for human intelligibility and/or
Make available to be program input
One UVC program for every file type (~1000 world-wide)
One Restore Application per future computer architecture

**Problem with proprietary formats**, e.g., MS Word files
Essential & Accidental Information

What’s essential depends on what somebody specific wants to accomplish (information producer or consumer). Accidental content cannot be avoided.

Imperfect knowledge capture is a fact of life; many examples in engineering specifications.
Durable Encoding for Computer Programs

One emulator for each current computer architecture
One UVC compiler or interpreter for each future computer architecture
Properties of TDO Methodology †

Format migration (which is error-prone) is not required
Consumers can evaluate authenticity without human administrative help
Metadata-to-object dissociation rare and detectable
Insensitive to Internet security risks
Links have message authentication codes for reliability
Implementable without disrupting existing repositories
Conformance with regulatory requirements will be easy
Implementation easy (based on XML and text editors)
Scalable because of recursion and heavy use of links

† Detailed description in forthcoming book by H.M. Gladney
Some Conclusions

Preservation of manufacturing data does not present unique technical challenges → look at what members of other disciplines are writing and saying.
But manufacturing does have specialized file formats and standards.

Divide and conquer → treat preservation as content mgmt. increment.

Focus on data structure first; later on data management.
Be sensitive to the teachings of scientific philosophy.
Age-old procedures for paper have digital counterparts.
Big challenge: persuading information providers to write metadata

We have an architecture, and are ready to start software design.
Backup Slides

To the man with a hammer, everything looks like a nail!
Trustworthy Digital Object (TDO)
Schema for a Digital Collection