Design for the Long Term: Authenticity and Object Representation

IS&T Archiving 2005

Adam Farquhar, The British Library
Agenda

■ Digital Object Management at the British Library
  ■ Drivers
  ■ DOM Programme
  ■ Key aspects of the DOM system

■ Focus on integrity and authenticity
  ■ Challenges
  ■ Our approach

■ Focus on object structure
  ■ Challenges
  ■ Our approach
The British Library has a duty of care to preserve non-print material in perpetuity
- Legal deposit legislation for non-print material
  - Royal assent in October 2003
  - Secondary legislation being established
- Existing voluntary deposit scheme operational since 2000

The British Library has extensive and growing digital assets
- Digitised versions of BL material from early ’90s onwards
  - Treasures, Collect Britain, Newspapers
- Electronic journals
- Sound Archive’s 15TB of material per year (with 50 year collection)
- New digitisation initiatives: newspapers, sound, theses, etc
- Collections from scientists and authors
- Web archiving
- Cartography and datasets, …

Plan for 500TB within five years
Digital Object Management (DOM) Programme

Our mission is to enable the United Kingdom to preserve and use its digital intellectual heritage forever.

Our vision is to create a management system for digital objects that will:
- Store and preserve any type of digital material in perpetuity
- Provide access to this material to users with appropriate permissions
- Ensure that the material is easy to find
- Ensure that users can view the material with contemporary applications
- Ensure that users can, where possible, experience material with the original look-and-feel
Key DOM System Design Features

- Disaster tolerant
  - Robust engineering
  - Multi-site design
  - Clear security boundaries
- Scaleable
  - 100s of Terabytes; millions of objects
- Ensure integrity and authenticity
- Flexible
  - Support multiple ingest and distribution services
  - Handle any kind of digital object
- Designed for the long term
- Cost effective
- OAIS Compatible
Integrity and Authenticity: Challenges

- Detect when a document is damaged
  - Paper has visible tears, folds, stains
- Detect when a document is modified
  - Paper shows visible signs of change
- Detect when a document is inserted after the fact
  - Shelfmark is written on binding
  - Paper is date-stamped on accession
  - Paper ages
- We need to give digital content the best properties of paper!
Integrity: Approach

- **Detect damage: Use checksum or digest**
  - Wide-spread approach; we use SHA-1
  - A digest is a (fixed length) summary or fingerprint
  - Change one bit in the object, and the digest changes a lot; it is **very hard** to create another object with the same digest

- **To check for damage**
  - Store the digest at ingest
  - Periodically recompute from bitstream, and test
  - Test will fail if digest or document are damaged

- **But what about a malicious operator who modifies the digest and the content?**
  - We lose. The archive is corrupted!
Authenticity: Approach

- **Detect modification: Digitally sign every object**
  - Digital signatures using public-key cryptography
  - I use my private key to encrypt the digest of an object
  - You use my public key to decrypt the digest and test it
  - As long as my private key is safe, no-one can forge my signature

- **But what about a malicious operator who has access to the private key?**
  - It is very hard to keep the private key safe in software

- **We use a specialised hardware solution**
  - The hardware is tamper-resistant and tamper-evident
  - Any changes require m-of-n people present
  - But anyone can verify the signature using standard software
Doing More with Digital Signatures

- A digital signature can sign any statement – not just the content
- We use the digital signature to permanently bind an identifier to each object
- The digital equivalent of writing the shelfmark on the binding
- We sign (approximately)
  \[
  \text{digest}(\text{identifier} + \text{digest(content)})
  \]

```
<domBoundObject>
  <contentDigest>2FA3CED4…</contentDigest>
  <domID>10001</domID>
</domBoundObject>
```
Authenticity: Approach

- Detect when a document is inserted
  - Extend the digital signature with a timestamp
- But what about a malicious operator who resets the system clock?
  - We lose!
- Our hardware solution provides trusted time-stamps
  - Well established technology use a time-stamp authority to get a trusted time
  - Protocols and methods are complex, but fairly easy to use
  - Cryptographic methods, audit trails, and a tamper-resistant internal hardware clock prevents spoofing
Layered Representation: Authenticity

- **Challenge:** Ensure technology independence
  - Software vendors may fail
  - New technologies, systems will provide better solutions

- **Approach:** Layered representation
  - Minimal assumptions: uniquely identified bytestreams
  - File system holds all important information
  - Every ingested content object exists in a bit-identical file
  - Authenticity information is stored in a separate file
  - Any secondary index or data can be built from the files

- **But isn't that inefficient?**
  - No – secondary indices allow efficient traversal, etc

- **But what if a malicious operator renamed the files?**
  - The authenticity file holds time-stamped signature
Layered Representation: Metadata

- **Challenge:** Associate metadata with content
- **Approach:**
  - The authenticity layer doesn't know about metadata
  - METS provides the framework for specifying metadata
    - Flexible and expressive
    - Supports descriptive, technical, administrative, and structural metadata
    - Supports multiple records of each type
  - The METS object links to the bitstream identifier
  - The authenticity layer ensures that the link can’t be modified
- **But won’t metadata standards evolve?**
  - The content bitstream and its signature remain untouched
Layered Representation: Relationships

- **Challenge:** Handle compound objects and other relationships
  - A serial issue has articles, pages
  - A newspaper issue has scanned pages, OCR text, articles that span pages

- **Approach:** Leverage METS
  - The next layers use METS to store named relations between objects

- **Challenge:** Update or withdraw objects without actually modifying them

- **Approach:** Ingest a new METS object with the appropriate relationships
  - Access layer sees only the latest generation
Conclusion

- The British Library is engaged in an ambitious programme to preserve the digital intellectual heritage of the United Kingdom.
- We are building a system that is:
  - Disaster tolerant, scaleable, flexible, cost effective.
- Today we focused on:
  - Ensuring the integrity and authenticity of digital material.
    - Hardware provides secure time-stamped signatures.
  - Establishing a flexible layered representation.

http://www.bl.uk/about/policies/dom/homepage.html