TIMBUS

TIMELESS BUSINESS

Sponsored by the European Commission Directorate for Information Society

Automation and Scalability in Digital Preservation

William Kilbride
william@dpconline.org

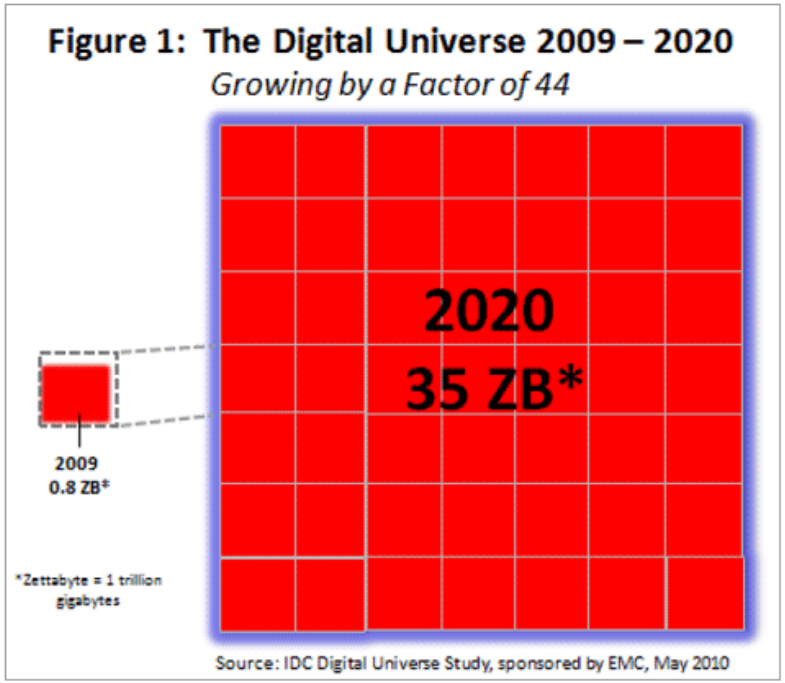
What's the problem?

Volumes of Digital Data are growing on (at least) three axes:

- Volume …
- Complexity …
- Importance …

'Digital Universe' Nears A Zettabyte

May 4th, 2010 : Rich Miller



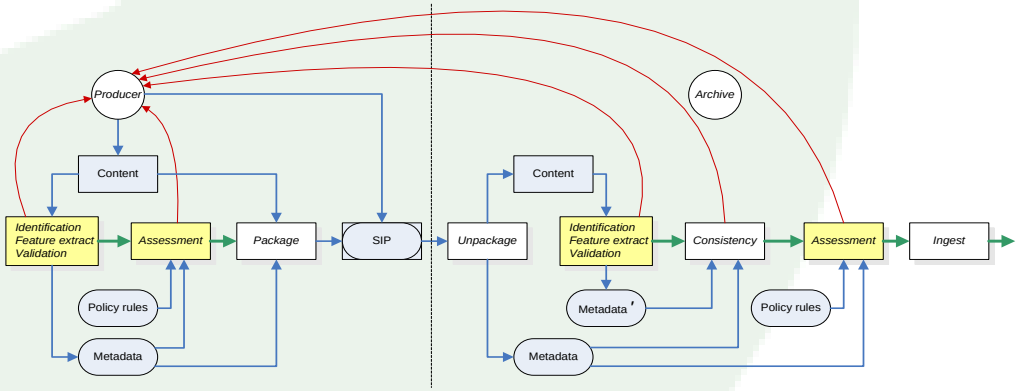
The Great Recession hasn't slowed the breakneck growth of the Digital Universe. In 2010 the volume of digital information created and duplicated in a year will reach 1.2 zettabytes, according to new data from IDC.

What's the problem?

Urgency to act – digital preservation intolerant of gaps

Budget ***and capacity*** not keeping pace with this growth:

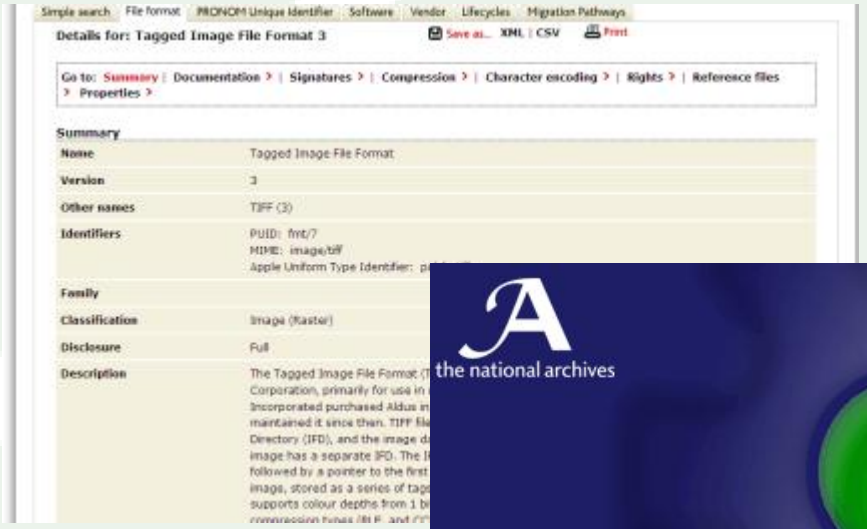
- Repetitious processing
- Bottlenecks, especially at ingest
- Manual intervention
- Staff costs
- Rapid development of tools



Automation 1: tools for characterisation

Making sense of what you've got at ingest:

- Pronom
- Droid
- JHove



- FITS
- UDFR (?)

Automation 2: tools for testing and management

Making one rational decision for many objects:

- DELOS Preservation Testbed
- LIFE 3
- PLATO 3

$$L_T = Aq + I_T + M_T + Ac_T + S_T + P_T$$

- PLATO 4
- APARSEN

Automation 3: tools for E-discovery

Forensics tools becoming more popular for ingest and management of large heterogeneous collections:

- FTK
- Sleuth Kit
- FIDO
- Bit Curator

Automation 4: automated workflows

Where possible, entire workflows from ingest to dissemination are being automated:

- ADS OASIS project
- UK Web Archive
- Digital continuity

Quality assurance is the key: getting the right kind of human intervention

Automation 5: automating access

Making digital preservation invisible to end users:

- Memento – time travel for the web
- Digital continuity

Making stuff easy to find therefore making value self evident

Automation 6: interoperability

Interoperability of preservation tools and services:

- PLANETS preservation interop framework
- CDL Micro-services framework

Currently harder to integrate interoperable services than to run them serially

Scalability and Automation: thoughts

1. Automated metadata extraction
   - *Natural language processing and sense-making is hard*
   - *Cultural versus technical preservation*
   - *Significant properties poorly defined*
2. Automated migration
   - *Too few migration tools and not enough tests*
   - *Migration 'on demand'*
3. End to end preservation
   - *Preservation tools not aligned*
   - *'Good enough is not always good enough'*