

Sponsored by the European Commission Directorate for Information Society

Automation and Scalability in Digital Preservation

William Kilbride william@dpconline.org



What's the problem?

TIMBUS

TIMBUS

TIMBUS

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

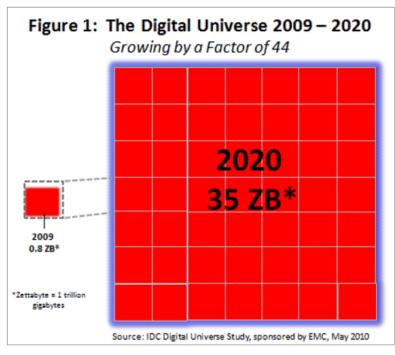
• Volume ...

TIMBUS

- 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?

TIMBUS

TIMBUS

Urgency to act – digital preservation intolerant of gaps

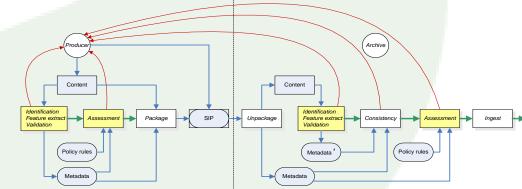
Budget *and capacity* not keeping pace with this growth:

Repetitious processing

TIMBUS

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:

Details for: Tage

Arailor

Identifiers

Family Classificatio Disclosure Description

9 Properties 9

- Pronom
- Droid

TIMBUS

JHove

- FITS
- UDFR (?)

vet ImopeOH Unique Identifier Software Vendor Lifecycles Migratian Pathways and Image File Formiat 3 🕑 Several XML CSV 🕮 Prent	
Documentation > Signatures > Compression > Character encoding > Wights > Reference files	
Tagged Image File Format	
3	
TIFF (3)	
PulD: fnt/7 MNR: image/bW Apple Unitorn Type Identifier: pr	
Image (faster)	
Rif	
The Tagged Image File Former () Corporation, primarily for use in Incorporated purchased Aldue in maintained it sins than. THE file Directory (IFD), and the image d image has a separate IFO. The II followed by a pointer to the first image, stored as a series of Tage supports colour depths from 1 bi convergedoon funda (IBLF) and CH	
Digital	
Record	2
©bject	Loading
Dentification	J

TIMBUS

TIMBUS



Automation 2: tools for testing and management

Making one rational decision for many objects:

- DELOS Preservation Testbed
- LIFE 3

TIMBUS

• PLATO 3

$\mathbf{L}_{\mathbf{T}} = \mathbf{A}\mathbf{q} + \mathbf{I}_{\mathbf{T}} + \mathbf{M}_{\mathbf{T}} + \mathbf{A}\mathbf{c}_{\mathbf{T}} + \mathbf{S}_{\mathbf{T}} + \mathbf{P}_{\mathbf{T}}$

- PLATO 4
- APARSEN



Automation 3: tools for E-discovery

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

• FTK

TIMBUS

- Sleuth Kit
- FIDO
- Bit Curator



TIMBUS IMBUS

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



TIMBUS

Automation 5: automating access

Making digital preservation invisible to end users:

C TIMBUS (TIMBUS) TIMBUS TIMBUS (TIMBUS) TIMBUS (TIMBUS) TIMBUS) TIMBUS

- Memento time travel for the web
- Digital continuity

Making stuff easy to find therefore making value self evident



TIMBUS

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



TIMBUS

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'



TIMBUS

TIMBUS