Safeguarding the Value and Ensuring the Longevity of Digital Assets

PNC 2013, Kyoto
10 December, 2013

Natasa Milic-Frayling
Principal Researcher
Microsoft Research
Physical artefacts

WYSIWYG Interface
Bravo. The first document preparation systems
1974 Butler Lampson and colleagues at Xerox PARC

Magna Carta Libertatum
Parchment document from 1297
Original document owned by David Farriero
Position

Digital is empowering but its properties must be understood in order to use it effectively.

How can we ensure that the contemporary digital content can be used in the future?
Overview

Digital Revolution and the Nature of Digital
Digital revolution and proliferation of electronic documents
Paradoxes of digital

Sustainability of Digital
Methods for ensuring content accessibility
Gaps in the value chain
Overview

Digital Revolution and the Nature of Digital

Digital revolution and proliferation of electronic documents
Paradoxes of digital
Digital revolution

Revolution led by advances in electrical engineering

- Information encoded in the digital form
- Computing instructions encoded in the digital form
- Electronic devices decode and process information, connect to other devices, and present information to humans to act on the computation results.
Mechanical computing

The Pascaline. Invented by Blaise Pascal (1623-62).

One of the first mechanical computing machines, around 1642.
Alan Turing had written the first programming manual for the Ferranti Mark I. His work on Mersenne Primes contributed to the extra orders added to the Ferranti Mark I, including the random number generator.

First commercially available general-purpose electronic computer
Digital computing

Early computing focused on computational problems in engineering, mathematics, sciences and, of course, writing love letters

From August 1953 to May 1954 strange love-letters appeared on the notice board of Manchester University’s Computer Department:
Digital computing

Early computing focused on computational problems in engineering, mathematics, sciences and, of course, writing love letters

From August 1953 to May 1954 strange love-letters appeared on the notice board of Manchester University’s Computer Department:

DARLING SWEETHEART
YOU ARE MY AVID FELLOW FEELING. MY AFFECTION CURIOUSLY CLING TO YOUR PASSIONATE WISH. MY LIKING YEARNS FOR YOUR HEART. YOU ARE MY WISTFUL SYMPATHY: MY TENDER LIKING.
YOURS BEAUTIFULLY
M. U. C.
From Mark 1 to contemporary computers

Alec Robinson at work on the Manchester Mark 1

http://www.computer50.org/mark1/gal1.html
Why is the digital revolution different?

Digital revolution advanced our core human capacity to innovate.

Digital computation enhances our abilities to conceive ideas, explore them, and translate them into action.
Re-inventing the science

- Use the computational power to collect observations and process the data
  - Mine the patterns
  - Create models from domain knowledge
  - Validate theoretical models and use them for prediction and control of the systems.

- Optimize dissemination of knowledge.

Both are pushing the boundaries of the current technologies and practices.
The Large Hadron Collider might find particles like the Higgs Boson — shown here as a simulation.

Photo: CERN
The Large Hadron Collider
Storing 15PB of data /yr \(^{(1)}\)(> 1.5TB / hr)

The LHC ATLAS sensor generates 1PB / sec \(^{(2)}\)

\(^{(1)}\)Databases for the Large Hadron Collector at CERN Dir Duellmann, CERN IT, XLDB Workshop, 2007
\(^{(2)}\)Marjorie Shapiro. (June 18, 2007). *Supersymmetry, Extra Dimensions and the Origin of Mass: Exploring the Nature of the Universe Using PetaScale Data Analysis*
Wiring and capacity of the human brain

• Decoding the complete “connectome” of the human brain is one of the great challenges of the 21st centuries.
  – Every second, billion of cortical nerve cells transmit billions of messages
  – 160 trillion synaptic connections
  – 300 million synopsis per cubic mm
  – Circuits that connect nerve cells are nanoscopic.
Evolution of computing infrastructure

- **Supercomputers**
  - Highly parallel, tightly synchronized message passing interface (MPI) simulations

- **Clusters**
  - Coarse grain parallelism, single administrative domains

- **Grids**
  - Job parallelism, throughput computing, heterogeneous administrative domains

- **Cloud**
  - Scalable, parallel, resilient web services.
Data Centers

- The Cloud is built on massive data centres (DCs)
- Size ranges from “edge” facilities to mega scale.

2009/2010: Approximate costs for a small size center (1K servers) and a larger (100K server) center:

<table>
<thead>
<tr>
<th>Technology</th>
<th>Cost in small Data Center</th>
<th>Cost in Large Data Center</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network</td>
<td>$95 per Mbps/month</td>
<td>$13 per Mbps/month</td>
<td>7.1</td>
</tr>
<tr>
<td>Storage</td>
<td>$2.20 per GB/month</td>
<td>$0.40 per GB/month</td>
<td>5.7</td>
</tr>
<tr>
<td>Administration</td>
<td>~140 servers/Administrator</td>
<td>&gt;1000 Servers/Administrator</td>
<td>7.1</td>
</tr>
</tbody>
</table>

Each data center is 11.5 times the size of a football field.
The benefits of Cloud

- Co-location of data and computation
- Economy of scale for hosting the services
- A framework to manage scalable, reliable, on-demand access to applications
- The "invisible" back-end to many of our computing devices and applications: the computational brain
- Prevalent business model "Pay as you go" access to "unlimited" remote data center capabilities.
Overview

Sustainability of Digital
Methods for ensuring content accessibility
Gaps in the value chain
Properties of digital

Enables computation at unprecedented speeds

Enables transfer of digitally encoded information at distances

Enables aggregation, analysis, and transformation of information.
From physical to digital

much larger than a rat-hole: she knelt down and looked along the passage into the loveliest garden you ever saw. How she longed to get out of that dark hall, and wander about among those beds of bright flowers and those cool fountains, but she could not even get her head through the doorway; “and even if my head would go through,” thought poor Alice, “it would be of very little use without my shoulders. Oh, how I wish I could shut up like a telescope! I think I could, if I only knew how to begin.” For, you see, so many out-of-the-way things had happened lately, that Alice had begun to think that very few things indeed were really impossible.

There seemed to be no use in waiting by the little door, so she went back to the table, half hoping she might find another key on it, or at any rate a book of rules for shutting people up like telescopes: this time she found a little bottle on it (“which certainly was not here before,” said Alice), and tied round the neck of the bottle was a paper label, with the words “DRINK ME” beautifully printed on it in large letters.

It was all very well to say “Drink me,” but the wise little Alice was not going to do that in a hurry. “No, I’ll look first,” she said, “and see whether it’s marked ‘poison’ or not,” for she had read several nice little stories about children who had got burnt, and eaten up by wild beasts, and other unpleasant things, all because they would not remember the simple rules their friends had taught them: such as, that a red-hot poker will burn you if you hold it too long: and that, if you cut your finger very deeply with a knife, it usually bleeds; and she had never forgotten that, if you drink much from a bottle marked “poison,” it is almost certain to disagree with you, sooner or later.
Increasing the utility

second time round, she came upon a low curtain she had not noticed before, and behind it was a little door about fifteen inches high: she tried the little golden key in the lock, and to her great delight it fitted!

Alice opened the door and found that it led into a small passage, not much

MSR prototype of flexible layout book reader from 1998 – Ralph Sommerer and Chuck Thucker
Born digital

US Government archive:

Major processing effort at the US National Archives when the Bush administration leaves office (2008).

The technical challenge is a result of the explosion in cyber communications

The electronic record of the Bush years about 50 times as large as that left by the Clinton White House in 2001.
but, it's not just size

“the Archives will also need to process reams of information locked in some quaint proprietary formats”.

– The record management system (RMS) index "consists of an implementation of a customized older version of Documentum, running on Oracle

– The photos are stored in a "proprietary photo management software called MerlinOne, running on Microsoft SQL as the database engine"

• It has taken several months to extract the images and metadata for re-linkage outside the Merlin format.
What is the problem?

Digital media is a victim of its own success: its rapid technological advances

- Document formats, software and hardware are becoming obsolete faster than we can ensure the forward compatibility of the content.
Bletchley Park, United Kingdom
Knowledge is in danger

Valuable data and knowledge resources are in the digital form, persisted in files.

Physical artefacts (manuscripts, books, audio and video tapes) are digitized to avoid perils of physical deterioration.

Yet, we do not have effective ways of preserving digital.
Paradoxes of digital

*Digital is persisted yet ephemeral*

*Without computing, digital is invisible and untouchable*
Digital needs a player

Digital encoding is inaccessible to human senses.

We cannot create digital without a digital tool.

We cannot sense digital without a software.

- Software interprets and presents the digital content into a human consumable form.
What is a digital object?

Digital Object = FILE?
Even contemporary file formats suffer from inconsistency of content rendering across the readers.

http://www.browsera.com/features#layoutProblems
Digital artefacts

• A digital document is ‘realized’ when a document file is processed through an application—a ‘reader’ or ‘viewer’

• Digital artefact is ephemeral—it exists only while the file is rendered and presented.

FILE – digital object

SOFTWARE – content decoder and generator

DIGITAL ARTEFACT

Persisted part of the digital artefact

DISPLAY—hardware and software for content rendering
Symbiosis of files and applications

FILE + APPLICATION → DIGITAL CONTENT

Persisted

Ephemeral
Computable digital artefacts

- Properties of the digital artefacts are determined by the software that produces the representation of the content.

- Digital artefacts can be fully generated computationally, e.g., fractal images, or can result from processing a digital file with data.
Paradox: we are concerned about storage, yet

Digital is inherently about *processing* bits, not about storing bits

*Storing bits is necessary but not sufficient to instantiate digital content*

Even when persisted and preserved, digital is *ephemeral*

*only through computation (rendering, playing, etc.) can we experience digital.*
Technical approaches to obsolescence

File Migration, Application Porting, Virtualization
Defining ‘digital preservation’

Objective of preservation is to ensure that the files are preserved and the software applications run in the contemporary computing ecosystem.
What do we want to keep ‘unchanged’?

- If application is not running in the contemporary environment
What do we want to keep ‘unchanged’?

- If application is not running in the contemporary environment
  - Migrate files and run with a contemporary software

(give up on both the original files and the application)
What do we want to keep ‘unchanged’?

- If application is not running in the contemporary environment
  - Migrate files and run with a contemporary software
  - Retain the files and port the application to the new environment
    (retain content files by give up on the application, at least partially)
What do we want to keep ‘unchanged’?

- If application is not running in the contemporary environment
  - Migrate files and run with a contemporary software
  - Retain the files and port the application to the new environment
  - Create a virtual machine with the old computing stack and run the original files and software.

(retain original files and original application; maintain scaffolding, i.e., VM environment)
Computational cradles

- VM solutions + Bridging Services
- Individual computational ‘cells’ for different generations of software stacks
- Bridging services: format translators, content extractors, etc.
Connecting legacy with contemporary ecosystem

Bridging Technologies and Methods

ICT SOFTWARE AND HARDWARE INNOVATION

It is all possible, of course. It is “only” software! We just need TIME and MONEY.
Economic sustainability

Gaps in the value chain
Software application lifespan

Applications rely upon a stack of software technologies. Maintaining an application is costly since the ecosystem is volatile. Any layer in the stack may change at any time. Software requires continuous upgrades to run within the existing ecosystem.
Digital media value chain

DIGITAL MEDIA HOLDERS AND KEEPERS

DIGITAL MEDIA CREATION AND CONSUMPTION

COMPUTING TECHNOLOGY INDUSTRY
Digital media value chain

**Computing Technology Industry**
- Tools and licenses for R&D cost
- Recovered $$

**Digital Media Consumption Creation and Digital Media Holders and Keepers**
- Value from produced content
- $$$$$$–$
- Future cost of sustaining digital media
- $(un)foreseen potential value in unspecified

### Unspecified
- **Recovered R&D cost**
- **Licenses for tools and services**
Digital media value chain - Risks

Risk of potential value loss (????) from evolution of the digital technologies

DIGITAL MEDIA HOLDERS AND KEEPERS

No value and no risk related to produced digital media. Driven to produce new technologies.

DIGITAL MEDIA CREATION AND CONSUMPTION

Risk of value loss ($–$$$$$$) across ICT development cycles

COMPUTING TECHNOLOGY INDUSTRY
Digital media value chain - Influence
<table>
<thead>
<tr>
<th>Professional Support</th>
<th>Business hours support for $259 U.S.</th>
<th>Business-critical After-hours support** for $515 U.S.</th>
<th>Order a 5-Pack Support Contract for $1,289 U.S.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Developers</td>
<td>Get Started</td>
<td>Get Started</td>
<td>Get Started</td>
</tr>
<tr>
<td>IT Professionals</td>
<td>Get Started</td>
<td>Get Started</td>
<td>Get Started</td>
</tr>
<tr>
<td>Partners (resellers/consultants)</td>
<td>(888) 456-5570</td>
<td>(888) 456-5570</td>
<td>(800) 936-3500</td>
</tr>
<tr>
<td>Microsoft Certified Partners</td>
<td>(888) 677-9444</td>
<td>(888) 677-9444</td>
<td>(800) 936-3500</td>
</tr>
<tr>
<td>Original Equipment Manufacturers</td>
<td>Get Started</td>
<td>Get Started</td>
<td>Get Started</td>
</tr>
<tr>
<td>System Builders</td>
<td>Get Started</td>
<td>Get Started</td>
<td>Get Started</td>
</tr>
</tbody>
</table>

*Note: ** indicates support for up to $1,289 U.S.*
Oracle Engineered Systems Price List
May 1, 2013

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exadata Storage Server X3-2 HP</td>
<td>55,000</td>
<td>6,600</td>
<td>4,400</td>
</tr>
<tr>
<td>Exadata Storage Server X3-2 HC</td>
<td>55,000</td>
<td>6,600</td>
<td>4,400</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exadata Database Machine X3-2 HP Full Rack</td>
<td>1,100,000</td>
<td>132,000</td>
<td>88,000</td>
<td>22,000</td>
<td>33,500</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HC Full Rack</td>
<td>1,100,000</td>
<td>132,000</td>
<td>88,000</td>
<td>22,000</td>
<td>33,500</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HP Half Rack</td>
<td>625,000</td>
<td>75,000</td>
<td>50,000</td>
<td>12,500</td>
<td>19,000</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HC Half Rack</td>
<td>625,000</td>
<td>75,000</td>
<td>50,000</td>
<td>12,500</td>
<td>19,000</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HP Quarter Rack</td>
<td>330,000</td>
<td>39,600</td>
<td>26,400</td>
<td>6,600</td>
<td>10,000</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HC Quarter Rack</td>
<td>330,000</td>
<td>39,600</td>
<td>26,400</td>
<td>6,600</td>
<td>10,000</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HP Eighth Rack</td>
<td>200,000</td>
<td>24,000</td>
<td>16,000</td>
<td>4,000</td>
<td>-</td>
</tr>
<tr>
<td>Exadata Database Machine X3-2 HC Eighth Rack</td>
<td>200,000</td>
<td>24,000</td>
<td>16,000</td>
<td>4,000</td>
<td>-</td>
</tr>
<tr>
<td>Exadata Database Machine X3-8 HP Full Rack</td>
<td>1,650,000</td>
<td>198,000</td>
<td>132,000</td>
<td>33,000</td>
<td>50,000</td>
</tr>
<tr>
<td>Exadata Database Machine X3-8 HC Full Rack</td>
<td>1,650,000</td>
<td>198,000</td>
<td>132,000</td>
<td>33,000</td>
<td>50,000</td>
</tr>
</tbody>
</table>
Supporting the digital lifecycle

requirements

Designing technologies with the view of the full life-cycle of tools, devices, and services

Business models that involve different distributions of wealth and risk associated with digital provisions for economically non-viable computation tools and services.
Cloud paradigm may enable digital future

*under the assumptions that:*

- Access to digital media becomes one of the primary drivers for innovation and evolution of the ICT ecosystem
  - Customers/digital media producers should demand and pay for long term access provisions at the time of technology acquisition.

- Digital media curation and education become an essential component of digital media services
  - Content creators and content holders need to demonstrate that there is value in combining contemporary and past information to provide compelling and competitive services.
Thank you

Natasa Milic-Frayling
natasamf@microsoft.com