Format Obsolescence and Validation

Making validation more than a proxy for rendering

Introduction

Your presenter

- Carl Wilson
- Technical Lead
- Open Preservation Foundation
- Email : carl@openpreservation.org
- Skype : carl.f.wilson
- GitHub : carlwilson
- Twitter : @openpreserve
- Google+ : carl@openpreservation.org

The Open Preservation Foundation

- International, not-for-profit membership organisation founded 2010
- Sustain the results of R&D projects
- Steward open-source digital preservation software
- Members & supporters are libraries, archives, universities, companies who provide digital preservation services

Todays talk

My plan:

- A quick history of physical obsolescence
- Examining obsolescence in the digital world
- The role of validation and its relationship to rendering

Obsolescence, know your enemy

What is Obsolescence?

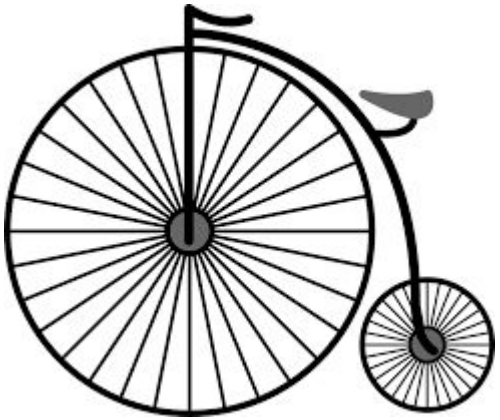
Latin root obsolēscere, which means "to fall into disuse."

- Obsolescence : the state of being which occurs when an object, service, or practice is no longer wanted even though it may still be in good working order.
  Wikipedia
- Obsolete: not in use any more, having been replaced by something newer and better or more fashionable.
  Cambridge Dictionary

It IS a popularity contest

Wikipedia also observes:

“Typically, obsolescence is preceded by a gradual decline in popularity.”



What causes obsolescence?

There are different types of obsolescence:

- Technical : something better comes along, it's evolution
- Functional : the item can no longer adequately perform its function, production or maintenance is no longer economically viable

But these are OK…..

Planned obsolescence

The centennial light bulb:

200 years and going strong.

Enter the Phoebus cartel

- GE, Phillips, Osram
- Bulb life managed down from 2,500 hours in 1924 to 1,000 hours in 1940

Planned obsolescence = good for the manufacturer

When nobody cares..

Style obsolescence is the lack of:

- support
- interest
- enthusiasm

And it’s hard to fight. Extreme example, who'd want to play in on deserted server, no matter how well the game’s preserved.

Me, myself and I..

Sustaining the unsustainable

- Something that is obsolete is almost by definition unsustainable.

BUT

- There are plenty examples of obsolete technology sustained by a community of care.

Perhaps there's a model for digital preservation here...

The obsolete in action

Preserving information

In the pre-digital world information was encoded onto a physical carrier.

- Ties the message to the medium
- Makes replication difficult
- If you want to preserve the information you must preserve the medium

Carrier/medium obsolescence

This means that it's possible for a carrier to become obsolete / disappear:

- wax cylinders
- old nitrocellulose film stock
- reel to reel tape

Casualties and survivors

- 80% - 90% of silent movies pre-1930 are considered lost due to unstable carrier. So the film archivists use digital now right?
- But carrier obsolescence can be survived

The digital killing zone

From the mid 80's there's been carrier carnage:

- cassette tape
- VHS vs. Betamax and gone
- LaserDisc
- Polaroids and Kodak film
- CD, SACD, DVD, HD-DVD, BLU-RAY??

Rendering obsolescence

It’s possible that carrier was OK e.g cassette tape but the technology to render became obsolete.

The story goes that an oil company had to build one of these in order to retrieve 1970’s level data samples.

Physical world obselescence

- Planned obsolescence is good for the manufacturer. Even media collections can become obsolete.
- Technological obsolescence means shiny new stuff and is good for the consumer?
- Both drive the consumer economy and it's a bit chicken and egg, do we have a built in love of obsolescence?

Digital Obsolescence Examined

Let’s get digital

Let’s examine obsolescence in the digital world while bearing in mind the physical world.

- It’s a revolution, so what’s changed?
- Is my life any easier?
- Where’s the problems?

Technical obsolescence

The new status quo, everything changes all of the time, but backwards compatibility is often easier to achieve in the digital world.

- When people care preservation takes place, e.g the emulation of old games

What happens when the fast moving world of technology collides with the conservative world of libraries and archives??

Carrier obsolescence

Separation of message from medium, or data from carrier means this can be avoided.

- Floppy disks, zip drives
- Variety of CD-RW / DVD-RW formats
- HDDs, SSDs, flash drives
- Variety of filesystems

Provide a multitude of potential baskets for your eggs.

Democratisation of production.

AND the means of replication

- While there's devices and software producing the format can live on.
- The separation of data and carrier makes replication trivial.
- There are established "popular" formats for images, music and video, documents, etc.
- An ecosystem of migration tools for interoperability production and processing software.

Redundancy of rendering.

The proliferation of devices and software capable of rendering digital content offers some security:

- PCs
- Tablets
- Mobiles
- Watches
- smart TVs
- eReaders

mean that rendering obsolescence may be less of an issue than we fear.

Proprietary problems

Proprietary efforts to protect content are an obsolescence risk:

- tie the data to a medium (protected discs) or the means of delivery (movie streaming).
- Badly / undocumented proprietary formats mean you don't own your content.
- proprietary rendering software means you're potentially one business decision away from obsolescence. Remember, these guys ARE capable of planning it.

Advantage open standards/source?

- Standardisation means detailed analysis of the standards and discussion to resolve ambiguity and yields robust formats
- Open and well documented standards mean that anyone can build their own parser, validator and renderer if necessary.
- Why not "own" the means to render, convert and produce your content?

The Role of Format Validation

OPF a vested interest in validators

What is format validation?

We talk about validators, what do they do

- parse the file according to a strict interpretation of the specification.
- fail if the file contravenes the specification.

Why do we validate?

In the belief that software based on the same specification will be able to:

- extract data from; or
- render the file

at a later date.

Foundations of faith

For this to work we have to trust in the universal language of standards specification which is:

- Formal language
- Wordy
- Designed by committee

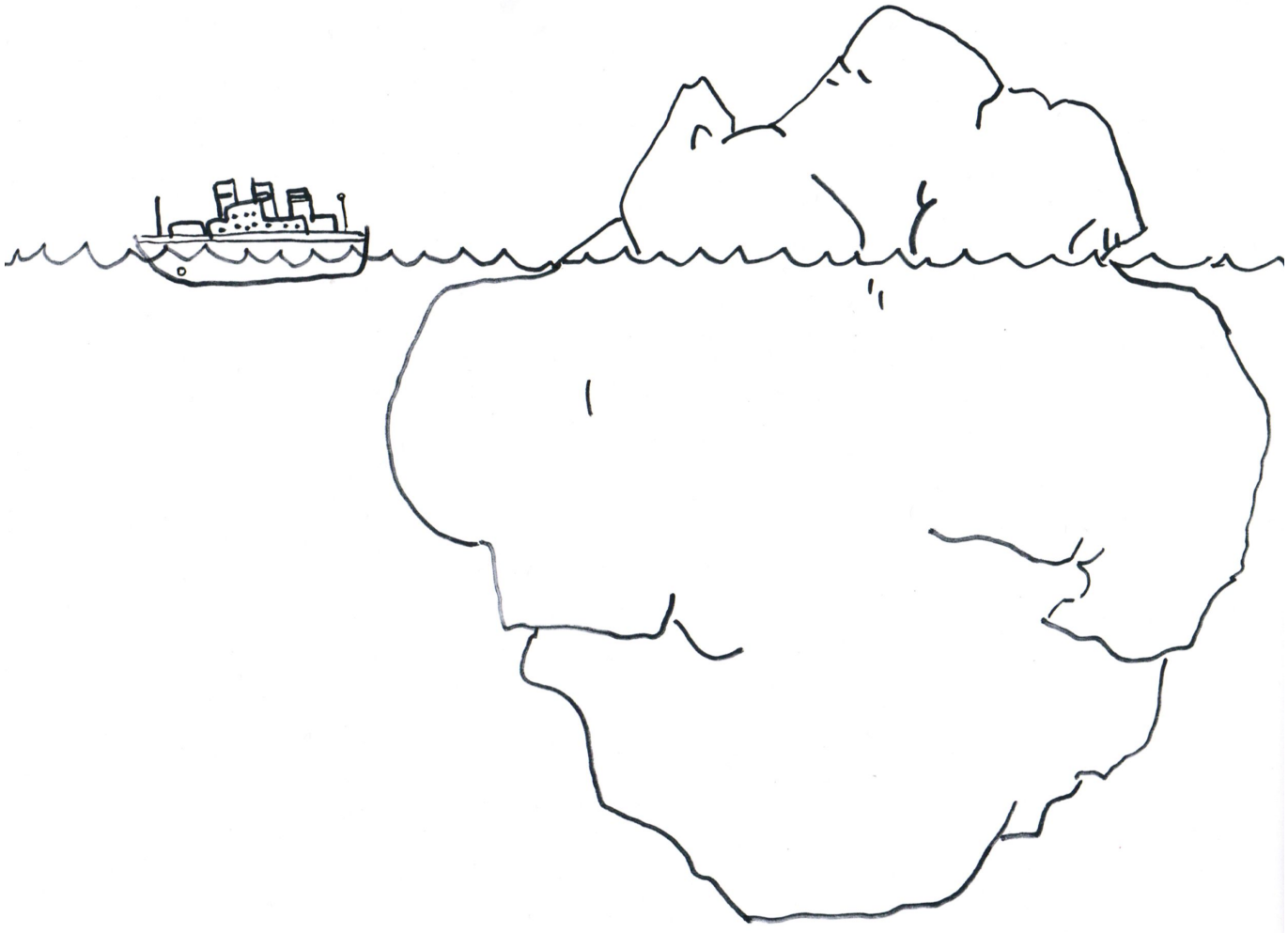
These documents are big, dry and difficult to read. Unsurprisingly not everyone interprets them in the same way.

PDF/A standards

PDF/A is a standard that comprises three parts:

- PDF/A-1
  ISO 19005-1:2005 Part 1: based on PDF 1.4
- PDF/A-2
  ISO 19005-2:2011 Part 2: based on ISO 32000-1 (PDF 1.7)
- PDF/A-3
  ISO 19005-3:2012 Part 3: based on ISO 32000-1 with support for embedded files

PDF/A is an iceberg



Open to (mis)interpretation

In some cases the specifications leave room for different interpretations of requirements. Such as:

- 6.1.6 Strings
  Hexadecimal strings shall contain an even number of non-white-space characters.
  Hexadecimal strings in PDF: <4E6F 7620 7368>

It is not clear, which characters are considered as white spaces. In particular, if a NULL (00h) character is allowed inside hexadecimal strings.

The gap between validation and rendering

PDF effectively runs two standards anyway:

- The strict standard enforced by validators
- The behaviour of the “conforming reader” which is far more forgiving

What your validators not telling you

- is the entire file valid, attachments, sub-formats, etc. might go unvalidated
- whether the file makes "sense" / is what you think / hope it is
- will you be able to render it in the future?
- if validation fails what are the REAL problems with the file and should I care?

Building a better, more valid world

What might improve matters?

- understanding the real preservation risk associated with validation errors.
- comprehensive test corpora suitable for testing validators and renderers.
- the development of open source combined validators / renderers that at least shared the same parsers and rules engine.