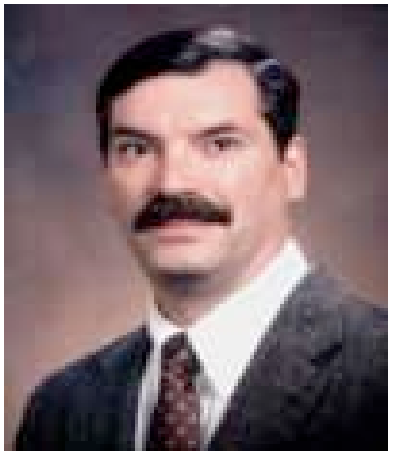
How are we “Ensuring the Longevity of Digital Documents”?

David S. H. Rosenthal

LOCKSS Program
Stanford University Libraries
http://www.lockss.org/
http://blog.dshr.org/

Kirk McKusick's IEEE Award



- 30 years of the Unix file system
  - Disks 1,000,000x bigger
  - Code 4x bigger, much faster, more reliable
- Reads every disk it ever wrote
  - No incompatible change to on-disk format
  - No incompatible change to API
- For widely used software
  - Costs of incompatibility outweigh benefits
  - *Strict compatibility makes Kirk's life easier*

Shifting Sands

"... digital documents are evolving so rapidly that shifts in the forms of documents must inevitably arise. New forms do not necessarily subsume their predecessors or provide compatibility with previous formats."

Jeff Rothenberg “Ensuring the Longevity of Digital Documents” *Scientific American* Vol. 272 No. 1 1995

As Jeff wrote this, Kirk's file system was 16 years old, with no incompatible changes to the API or on-disk format.

The Meme

- Incompatibility is inevitable, a force of nature
  - *Why did Jeff think this in 1995?*
  - *Is it true in 2009?*
- If this meme isn't true
  - *What causes incompatibility?*
  - *Are these causes operating now?*

Talk in 3 Parts

- Ancient History: before 1995
  - Jeff Rothenberg's 50-year look forward from 1995
  - What he predicted & why
- Modern History: from 1995 to 2009
  - Impacts of Jeff's article
  - What else happened
  - How Jeff rates as a prophet & why
- The Future: following Jeff's example
  - Looking forward to identify the real problems

Ancient History

- "History is not what you thought. *It is what you can remember.* All other history defeats itself."
  - From the *Compulsory Preface* to *1066 And All That*, W. C. Sellar & R. J. Yeatman

Jeff Rothenberg's Scenario

- In 2045, descendants find a CD
  - Try to recover document from it leading to Jeff's fortune
- Threat: Media degradation
  - Bits on the CD suffer “bit rot”
- Threat: Media obsolescence
  - No hardware capable of reading the bits available
- Threat: Format obsolescence
  - No software capable of rendering the bits available

Jeff on Format Obsolescence

- Defenses
  - Format Migration
  - Emulation
- Format migration disapproved
  - “Finally, [format migration] suffers from a fatal flaw. ... Shifts of this kind make it difficult or impossible to translate old documents into new standard forms.”
- Emulation approved subject to caveat
  - “specifications for the outdated hardware ... must be saved in a digital form independent of ... software”

Jeff's Dystopian Vision

- Documents survive in off-line media
- The media have a short lifetime
- The media readers have a short lifetime
- Documents are in app-specific formats
  - Typical formats are proprietary
  - Attempts to standardize formats will fail
- Hardware & O/S will change rapidly
  - In ways that break applications
- Apps for rendering formats have a short life

Two Words: Desktop Publishing

- The publishing medium was paper
- Design goal of *Word* & *WordPerfect* files:
  - Save the state of the word processor
- Formats - exclusive property of applications
  - Other apps interpreting them – threat to biz model
- Then people started e-mailing the files:
  - Got there quicker, could be edited & returned

IT in 1995

- Many hardware architectures
  - X86, SPARC, MIPS, 680X0, PowerPC, ...
  - PC split between ISA bus and PCI bus
- Several operating systems
  - Windows 3.X, Windows 95, OS/2, System 7, Solaris, ...
  - Linux (1.2.0) was barely functional
- Fragmented applications market
  - MSFT Word vs. WordPerfect ...
  - Lotus 123 vs. Excel
  - No standard for PC graphics, so no 3D PC games

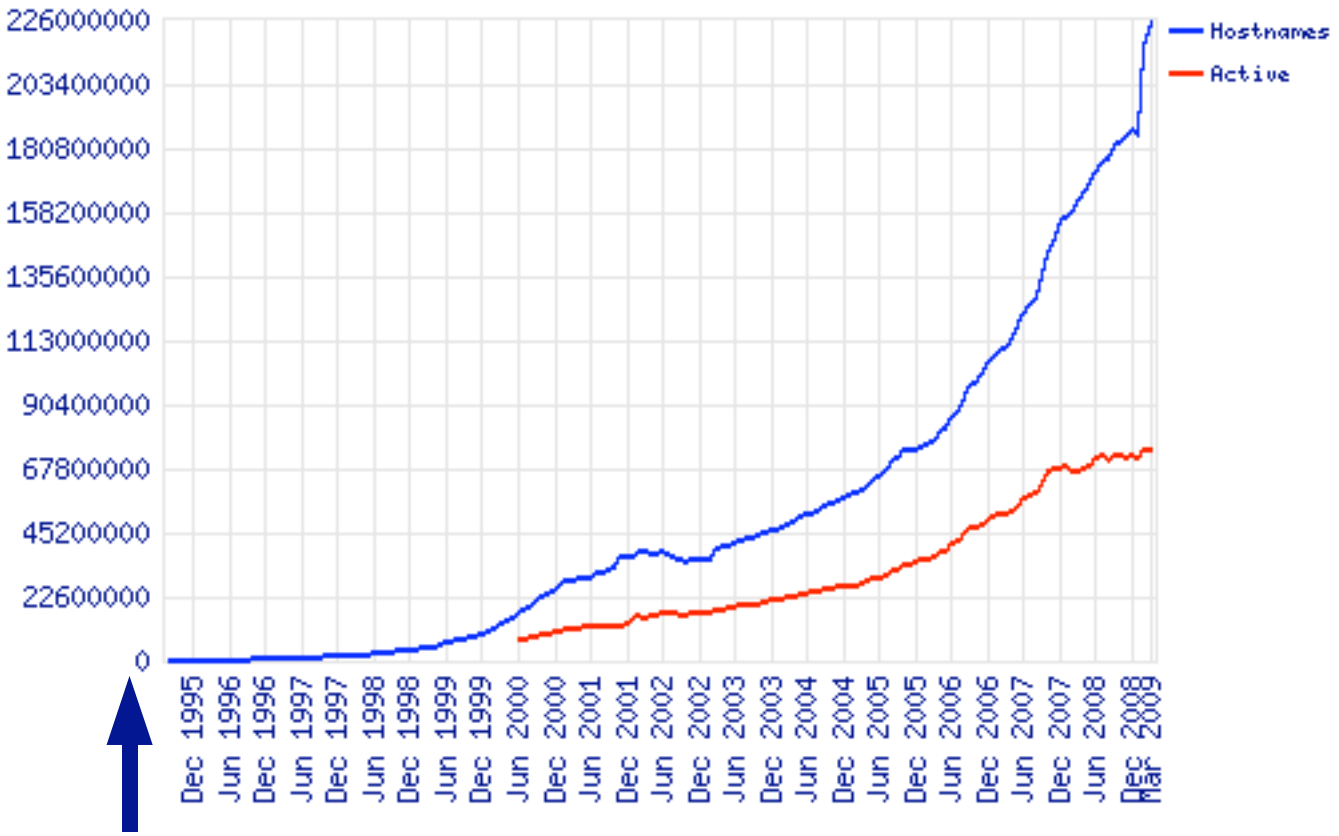
Modern History

- “A preoccupation with the future not only prevents us from seeing the present as it is but often prompts us to rearrange the past.”
  - Eric Hoffer

Impacts of Jeff's Vision

- Scientific American article = lots of attention
- Governments, foundations started funding
  - Mellon Foundation
  - NSF, Library of Congress, National Archives ...
- Now have systems in production
  - Using both strategies Jeff identified
- Internet Archive started the next year
  - Using neither of them

The Web!



- May 1995: HighWire puts *JBC* on-line
  - Pioneers academic e-journals

Off-line or On-line?

- In Jeff's vision documents survived off-line
  - Coming on-line for occasional manipulation or copying
  - Copy-ability was *extrinsic* to the medium
- Now, if it is worth keeping, it is on-line
  - Off-line backups are temporary
- Copy-ability is *intrinsic* to the on-line medium
  - No-one cares what the physical medium is
    - Disk, flash memory, RAM, ...
  - Just that it obeys the access protocols

Microsoft vs. its Users

- MSFT Office biz model has to drive upgrades
  - Introduce gratuitous format incompatibility by default
  - New machine writes document old machine can't read
  - Old machine buys upgrade, MSFT happy
- Users carry the cost of incompatibility
  - Unhappy – anti-trust probe ('90) & consent decree ('94)
  - Users ('02-'05) force ODF standard for documents
  - MSFT ('07) does OOXML, but concedes the basic point
- Experience with MSFT misled Jeff
  - Even MSFT's ability to obsolete formats now limited

Documents or Content?

- Jeff's documents were property of a program
  - A *Word* file is data to be manipulated (only) by *Word*
  - Proprietary format changeable on a whim
- Now documents are content to be published
  - Charge to upgrade browser so it can't read old content?
    - Browser free, content free, *Office* biz model dead
  - Goal of publishing: reach as many readers as you can
  - Gratuitous incompatibility is now self-defeating
    - Publishing *IE*-only pages gets you flamed

Virtual Machines

- H/W virtualization has long history (VM/370!)
  - Software too (Basic!)
- In 1995 it wasn't mainstream
  - Intel was just putting necessary stuff into X86
- Now virtual hardware is mainstream
  - Old hardware can be emulated easily with open source
- Mainstream software now written for VMs
  - Java, C#, ...
- Jeff was right about emulation
  - But preservation wasn't the reason for doing it

Open Source

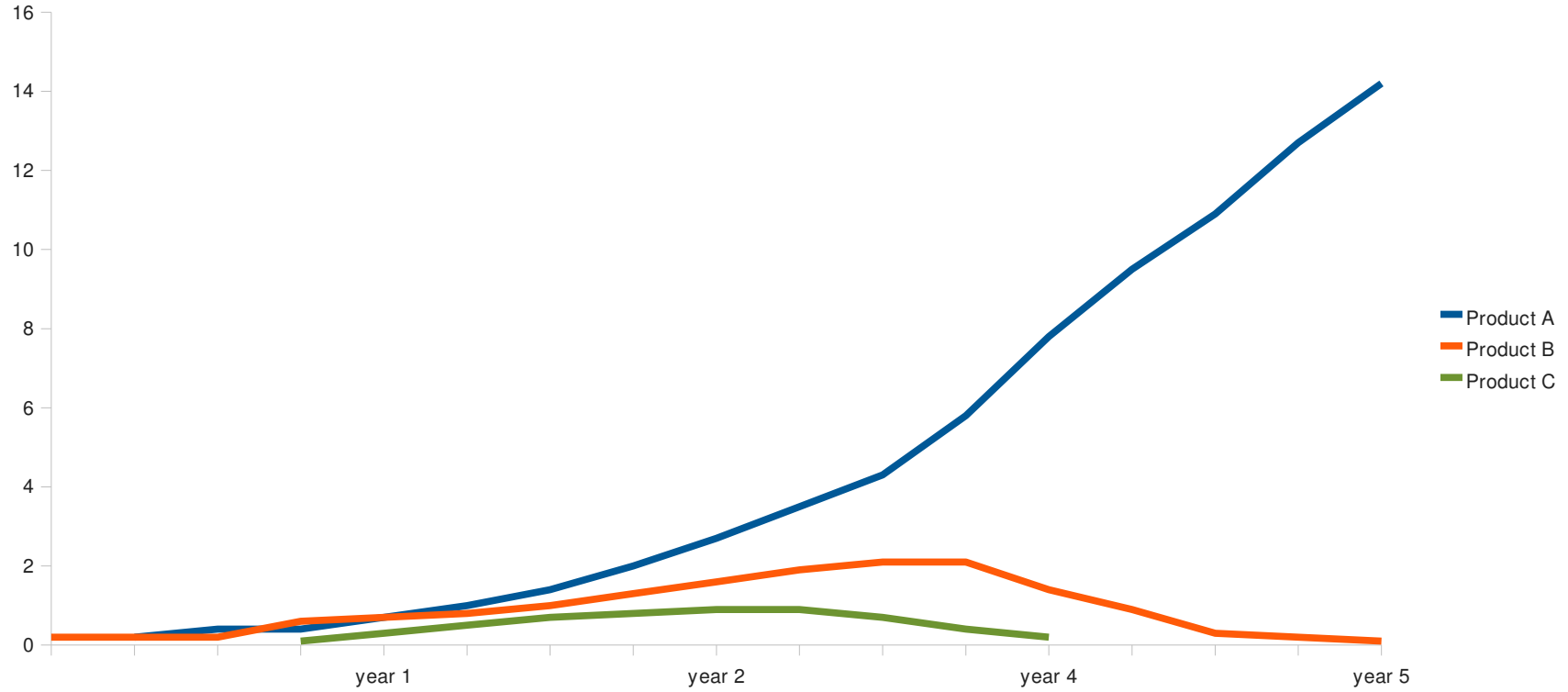
- In 1995 Open Source wasn't mainstream
  - Now it's basic strategy for all but 2 big IT companies
- Open Source renderers for all major formats
  - Even those with DRM! (Legal status obscure)
- Open Source is best preserved of all
  - ASCII, source code control, can rebuild stack as it was
- Open Source isn't backwards incompatible
  - For same reason as “no flag day on the Internet”
- Format with Open Source renderer is safe
  - Executable “preservation metadata”

20/20 Hindsight

- Documents survive on-line, on the Web
  - Off-line used only for temporary backups
- Migration between on-line media is inherent
  - Readers are bundled with storage technology
- Formats are standard & app-independent
  - Proprietary formats get open-source renderers
- Format obsolescence never happens
  - No flag day on the Internet
- I.e: Jeff wrong in every particular

The Big Picture

- *Increasing Returns & Path Dependence in the Economy* W. Brian Arthur (1994)

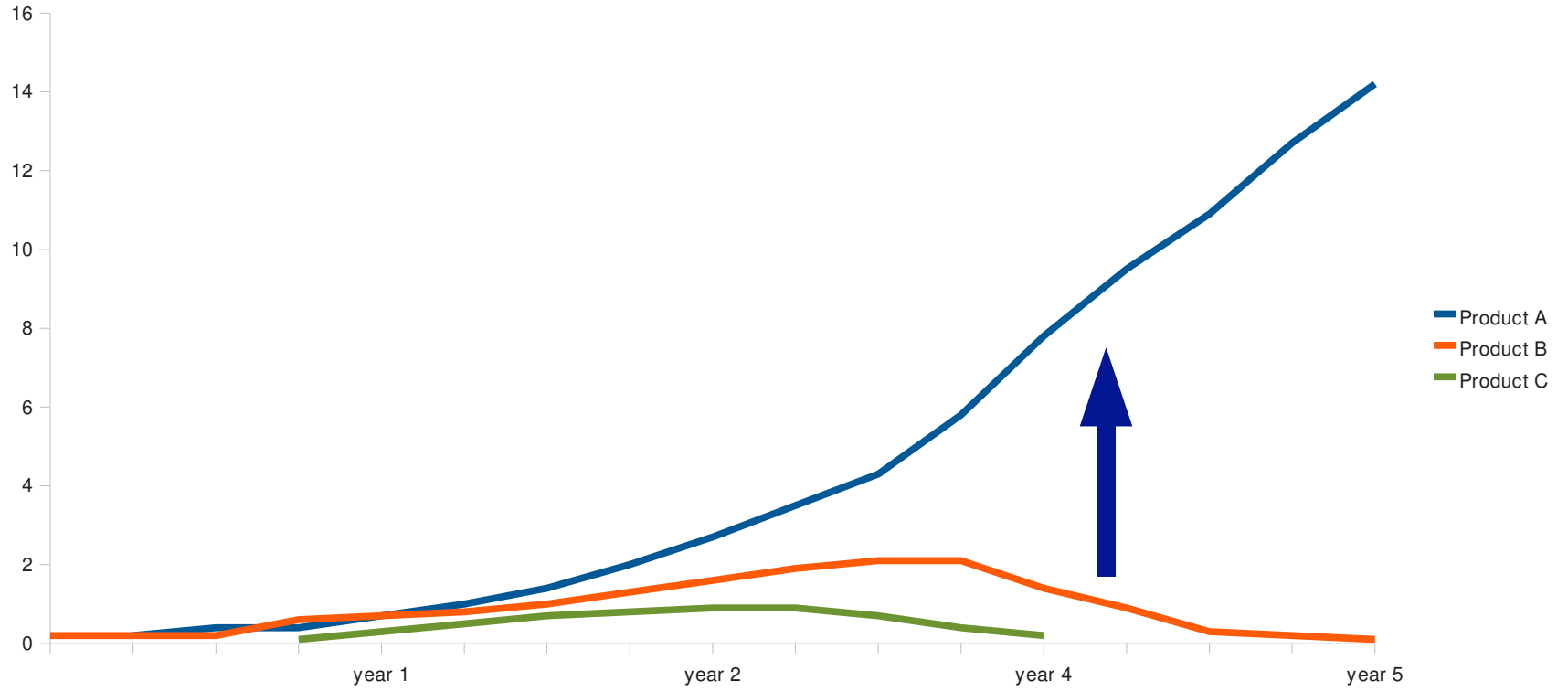


The Big Picture

- IT markets have increasing returns
  - Usually called “network effects” - Metcalfe's Law
- IT markets have path dependence
  - Many players early
  - Randomly one gets bigger, network effects take over
- IT markets subject to capture (MSFT, INTC)
  - Captured markets slow change down (e.g. Vista)
- History misled Jeff to overestimate change

The Big Picture

- Users migrate to the winner
  - Winner motivated to make this easy



Yes We Can!

- Jeff being wrong is Good News!
  - Collections that survive aren't as hard as we thought
- Just collect and keep the bits
  - Not collecting is the major reason for stuff being lost
- If you keep the bits, all will be well
  - Current tools will let you access them for a long time
- Just go do it!

The Future

- “Prediction is very difficult, especially about the future.”
  - Neils Bohr

The Real Problems Were ...

- Scale
  - Not individual documents but vast collections of them
- Cost
  - Preservation not by individuals but large organizations
- Intellectual Property
  - If content worth saving someone is making money from it

Scale

- Jeff looked at micro-level preservation
  - A single document on a single CD
- Society needs macro-level preservation
  - Information is now industrial scale
  - Data centers the size of car factories
  - As much power as an aluminum smelter
- 1 copy of 1 important database = $1M/yr
  - In storage costs alone
- Document-at-a-time preservation impractical
  - Curators must get huge collections per day's work

Metcalfe's Law

- The lesson of Google
  - More value in *connections* than in documents themselves
  - Preserving individual documents loses this value
  - Need to preserve collections *including* the connections
- Another instance of Metcalfe's Law
  - Value of a network goes as # of nodes squared
  - Isolated document is a network of 1 node
- Google's other lesson – it's expensive
  - We lack good cost data for digital preservation at scale
  - Use two extremes to get a ballpark estimate

Scale Implies Cost

- Internet Archive:
  - contains 2PB, growing 240TB/yr
  - Google collects the Web monthly then *discards* it
  - archive.org collects the Web monthly then *keeps* it
  - 2 snapshot copies + 1 coming up
  - $10-14M/yr operation so ~$0.5 per GB per year
- Portico:
  - All academic literature ~50TB, growing ~5TB/yr
  - Portico still working on ingesting back content
  - $6-8M/yr operation so >$10 per GB per year

How Many $ Do We Need?

- archive.org should be cheaper than Portico
  - It isn't doing all that “preservation” stuff
  - Better bit preservation than archive.org important
  - But does all the other stuff justify 20x cost per byte?
- How much do we need to save? An exabyte?
  - 0.3% of the data generated in 2007, 0.05% of 2011
  - @ archive.org = $5B/yr, @ Portico = $100B/yr
  - The world doesn't have even $5B/yr to spend on this

Intellectual Property

- Most content worth saving is making money
  - Lawyers won't risk that; don't want you to keep a copy
- They have massaged the law to their ends
  - You must get permission, so you must talk to lawyers
    - Or you are vulnerable to DMCA take-down like IA
- 1 hour of 1 lawyer ≈ 5TB of disk
  - 10 hours of 1 lawyer could store the academic literature
- For preservation, much uncertainty
  - Effort devoted to high byte/lawyer-hour content
- *Please* use Creative Commons licenses!

Looking Forwards

- What are the non-problems?
  - Or rather, the problems not big enough to matter
- What are the big problems?
  - Preserving the world the way it is now
    - Not the way it used to be
  - Finding enough money
    - And working out how much that is
  - Surviving not having enough money
    - By turning more things into non-problems

Non-Problems

- Formats
  - Any format with an open-source renderer is not at risk
- Metadata (at least for documents)
  - Hand-generated metadata
    - Too expensive, search is better & more up-to-date
  - Program-generated metadata
    - Why save the output? You can save the program!

Services not Documents

- "Preservation" implies static, isolated object
  - Web 0.9 is like reading a printed book
  - Web 1.0 dynamically inserts personalized adverts
  - No-one preserves the adverts, but they're important
    - *With the Night Mail* Rudyard Kipling (1905)
    - *The Who Sell Out* The Who (1967)
    - *A Prairie Home Companion* Garrison Keillor (1974-)
- Web 2.0 is dynamic, interconnected
  - Each page view is unique, mash-ed up from services
  - Pages change as you watch them
  - *What does it mean to preserve a unique, dynamic page?*

Things Worth Preserving

- User Generated Content
  - To understand 2008 election you need to save blogs
  - To do that you need to save YouTube, photo sites, ...
    - So that the links to them keep working ...
  - Technical, legal, scale obstacles almost insuperable
- Multi-player games & virtual worlds
  - Even if you could get the data and invest in the servers
  - They're dead without the community – *Myst* (1993)
- Dynamic databases & links to them
  - e.g. Google Earth mash-ups – is Google Earth forever?

Economics

- 2008 Preservation Buzzword: *Sustainability*
  - We can't afford to preserve the stuff we know how to
- Future stuff will be much more expensive
  - There'll be a lot more bytes of it
  - Each byte will be more difficult & more expensive
- Bytes vulnerable to money supply glitches
  - Data needs to be *endowed* if it is to survive hard times
  - Endowing up front means preserving less
- Collection development: what must be kept?
  - But it has really bad scaling problems

Digital Preservation Difficult

- Conceptually
  - What does it mean to preserve dynamic content?
- Technically
  - Need to preserve *services* not content. How?
- Legally
  - Preservation requires permission
  - How do you even find everyone you need to ask?
- Economically
  - Just storing the bits needs industrial infrastructure
    - Beyond resources of universities, national libraries
  - Are services like S3 reliable enough?

Digital Preservation Important

- Paper's attributes built in to society
  - Durable, write-once, tamper-evident, highly replicated, ...
- Society needs fixed, tamper-evident record
  - E.g. laws, contracts, evidence, ...
  - Paper provides this as a side-effect
- The Web is Winston Smith's dream machine
  - All govt. information on a single web server (FDsys)
  - Point-&-click to rewrite history

Practical Next Steps

- Everyone – just go collect the bits:
  - Not hard or costly to do a good enough job
  - *Please* use Creative Commons licenses
- Preserve Open Source repositories:
  - Easy & vital: no legal, technical or scale barriers
- Support Open Source renderers & emulators
- Support research into preservation tech:
  - How to preserve bits adequately & affordably?
  - How to preserve this decade's dynamic web of services?
    - Not just last decade's static web of pages