From shoeboxes to software preservation

Slides: https://doi.org/10.6084/m9.figshare.8088290
7th May 2019, Insert Coin to Continue: DPC Briefing Day on Software Preservation
Neil Chue Hong (@npch), Software Sustainability Institute
ORCID: 0000-0002-8876-7606 | N.ChueHong@software.ac.uk

Supported by:
Andy Warhol’s Time Capsules

FINDING AID FOR TIME CAPSULE 566
TIME CAPSULES CATALOGUING PROJECT
THE ANDY WARHOL MUSEUM ARCHIVES

DESCRIPTIVE SUMMARY

TITLE: TC566
REPOSITORY: The Andy Warhol Museum, 117 Sandusky Street, Pittsburgh, PA 15212
EXTENT: 1 linear foot (1 box), 24 x 15 x 10 inches, 231 objects (897 bulk objects)
CATALOGUED BY: TC Project Cataloguer Marie Elia
CREATOR: Warhol, Andy, 1928-1987

ARRANGEMENT SUMMARY

Time Capsule 566 contains 16 series with additional subseries arranged alphabetically. Series include Audio Material, Books, Business/Financial Records, Clippings, Correspondence, Costume/Personal Accessories, Ephemera, Equipment/Tools/Materials, Exhibition Announcements, Food/Products, Furnishings/Textile, Invitations, Manuscript Material, Photographic Material, Printed Material, and Serials. Unless otherwise noted, the arrangement scheme for the collection was imposed during processing in the absence of a usable original order.

ADMINISTRATIVE INFORMATION

Provenance: Donated to The Andy Warhol Museum by The Andy Warhol Foundation for the Visual Arts; consult museum archives for additional details
Restrictions: Access may be restricted; consult repository for details
Copyright: Copyright queries should be directed to the Rights and Reproductions Division of the The Andy Warhol Museum
Credit Line: The Andy Warhol Museum, Pittsburgh, Founding Collection, Contribution The Andy Warhol Foundation for the Visual Arts, Inc.

All photos from The Andy Warhol Museum apart from picture of Andy Warhol (Public Domain via Wikipedia)
Digital Domesday?

Domesday page – Warwickshire (public domain)
Domesday équipement By Regregex - Own work, CC BY 3.0,
https://commons.wikimedia.org/w/index.php?curid=10716074
Domesday Reloaded

What is Domesday?

In 1986 the BBC launched an ambitious project to record a snapshot of everyday life across the UK for future generations. A million volunteers took part?

Now, 25 years later you can explore the archive online, see the pictures, update the information and make your mark on this fascinating record of our collective history.

Read more about the story of Domesday here

Search the Domesday Site

https://webarchive.nationalarchives.gov.uk/20110911075344tf_/http://www.bbc.co.uk/history/domesday
Archiving Models

- Time between CAD Versions: 6 months
- Life of CAD System: 10 years
- Life of Product: 70 years +

Production, Services, Spares, Modifications

CAD Obsolete, CAD Forgotten, Legal Liability

10 years, 20, 30, 40, 50, 60

Image courtesy PDES Inc
Slide from Sean Barker, BAE SYSTEMS, DPC Designed to Last
https://slideplayer.com/slide/10521357/
They are
by Neil P Chue Hong

- internship applications now open!
Year - Fall Meeting, here I come!
Reinforcement system make it happen
To meet them. Hope they get home.

And thought - means business!
Did some work on this in the past.
But that's a pretty manual process
Of forming cohorts that last

Research and numerical simulations
“Success with the Pole” instead?
Researchers and institutions
Be plenty sleepless nights ahead
Remotely)? Here’s the instructions:

https://poetweet.com.br/
Experiential Culture

Image: thatgamecompany/Sony Computer Entertainment
Email Archives

Images from The Good Life (Enron Simulator): https://enron.email/

Re: Dark Star

Richard B <everyone@enron.email>

To: me

After further thought, it seems to me that in light of our fear of litigation with American Coal, we should keep the documents. To further insulate the Coal Group and you from any claim that Enron misused the information, I suggest that you transfer the information to me and I will hold it for safekeeping.

Good morning!

Mark - BCT Legal <everyone@enron.email>

To: me

Hope you’re having a pleasant first week of 1999. Thought I would forward this on...I found that 18, 19, 15, 14, 8, 7, 2 and I hit a little too close to home.

TOP 22 SIGNS THAT YOU HAVE HAD TOO MUCH OF THE 90s

18. Cleaning up the dining area means getting the last food bags out of the back seat of your car.
19. Your reason for not storing your sundries with family is that they do not have e-mail addresses.
20. Keeping up with sports emails and ESPN’s home page to your bookmarks.
21. Your list of “to do” list that includes entries for lunch and bathroom breaks, and they are the ones that never get crossed off.
22. You have actually finished your Christmas list to your parents.
23. Pick-up lines now include a reference to liquid assets and capital gains.
24. You consider second-day air delivery painfully slow.
25. You assume the question “so valet-park or not?” is rhetorical.
26. You refer to your dining-room table as the flat filing cabinet.

Images from The Good Life (Enron Simulator): https://enron.email/

Enron Email Dataset: https://www.cs.cmu.edu/~./enron/
ePADD: https://epadd.stanford.edu/
Social Movements

Photo by Sticker You on Unsplash

Photo by Mihai Surdu on Unsplash
The Need for Openness in Data Journalism

Brian Keegan, Ph.D. (@bkeegan) College of Humanities and Social Sciences, Northeastern University

Do films that pass the Bechdel Test make more money for their producers? I’ve replicated Walt Hickey’s recent article in FiveThirtyEight to find out. My results confirm his own in part, but also find notable differences that point the need for clarification at a minimum. While I am far from the first to make this argument, this case is illustrative of a larger need for journalism and other data-driven enterprises to borrow from hard-won scientific practices of sharing data and code as well as supporting the review and revision of findings. This admittedly lengthy post is a criticism of not only this particular case but also an attempt to work through what open data journalism could look like.

The Angle: Data Journalism should emulate the openness of science

New data-driven journalists such as FiveThirtyEight have faced criticism from many quarters and the critiques, particularly around the naïveté of assuming credentialed experts can be bowled over by quantitative analysis so easily as the terrifyingly innumerate pundit who infest our political media [1,2,3,4]. While I find these critiques persuasive, I depart from them here to instead argue that I have found this “new” brand of data journalism disappointing foremost because it wants to perform science without abiding by scientific norms.

The questions of demystifying what is or is not science are fraught, so let’s instead label my gripe a “failure to be open.” By openness, I don’t mean users commenting on articles or publishing whistleblowers’ documents. I mean “openness” more in the sense of “open source

A different model

The model above inexplicably uses “budget” on both sides of the equation, which is a no-no. However, we constructed R2 as \((Revenue - Budget) / Budget \) so in these models budget ends up being a function of itself.

What happens if we just leave Budget on the right side of the equation and simply estimate Revenue as a function of Bechdel rating and controlling for Budget?

We get very different findings. Now there is a significant and positive relationship between Budget and Revenue, as we’d expect. Furthermore, there is also a significant and positive relationship between Bechdel rating and Revenue. Better scores for women translates into better revenue, even controlling for the fact that bigger budgets also accrue more revenue. This model also explains approximately 28% of the variance, versus 10% in the article’s model suggesting it’s doing a better job modeling the relationship of Bechdel test and financial performance.

# Attempt to model

```python
model = smf.ols(formula='log(Revenue) ~ C(rating) + log(Budget) + data', data=adj_revenue)
model_full = smf.ols(formula='Revenue ~ C(rating) + log(Budget) + data', data=adj_revenue)
```

## Old Regression Results

<table>
<thead>
<tr>
<th>Dep. Variable</th>
<th>log(Adj_Revenue+1)</th>
<th>R-squared</th>
<th>adj. R-squared</th>
<th>Method</th>
<th>Least Squares</th>
<th>F-statistic</th>
<th>Prob (F-statistic)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data</td>
<td>09/19/20</td>
<td>0.320</td>
<td>0.294</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
<tr>
<td>Time</td>
<td>03/19/20</td>
<td>0.330</td>
<td>0.300</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
<tr>
<td>No Observations</td>
<td>1035</td>
<td>0.469</td>
<td>0.469</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
<tr>
<td>No. Residuals</td>
<td>1578</td>
<td>0.250</td>
<td>0.250</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
<tr>
<td>DF Residuals</td>
<td>1578</td>
<td>0.250</td>
<td>0.250</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
<tr>
<td>DF Model</td>
<td>1577</td>
<td>0.250</td>
<td>0.250</td>
<td>0.024</td>
<td>0.224</td>
<td>1.272</td>
<td>0.232</td>
</tr>
</tbody>
</table>

---

| Intercept     | 1.367               | 0.099    | 2.350          | 1.367  | 0.099         | 2.350       | 1.367             |
| Rating[1.5-4] | 0.226               | 0.032    | 2.350          | 0.226  | 0.032         | 2.350       | 0.226             |
| Rating[2.5-4] | 0.175               | 0.029    | 2.350          | 0.175  | 0.029         | 2.350       | 0.175             |
| Rating[3.5-4] | 0.125               | 0.026    | 2.350          | 0.125  | 0.026         | 2.350       | 0.125             |
| Rating[4.5-4] | 0.072               | 0.022    | 2.350          | 0.072  | 0.022         | 2.350       | 0.072             |
| Rating[5-4]   | 0.022               | 0.022    | 2.350          | 0.022  | 0.022         | 2.350       | 0.022             |
| log(Budget)   | 0.052               | 0.052    | 2.350          | 0.052  | 0.052         | 2.350       | 0.052             |

---

| Omnibus       | 1491.278            | 11.999   | 111.999        | 111.999| 11.999        | 111.999     | 111.999         |
| Freq(Omnibus) | 0.000               | 0.000    | 0.000          | 0.000  | 0.000         | 0.000       | 0.000           |
| Signif(Omnibus) | P(>F) | 0.000    | 0.000          | 0.000  | 0.000         | 0.000       | 0.000           |

---

Brian C. Keegan “The Need for Openness in Data Journalism”
https://nbviewer.jupyter.org/github/briankeegan/Bechdel/blob/master/Bechdel_test.ipynb
“Lack of access to effective software continues to be a major hindrance to scientific progress and therapeutic discovery. [...] For the benefit of all society, we need to pursue new and complementary approaches to the creation and dissemination of scientific software.”

– Warren Lyford DeLano, creator of PyMOL, in 2003
How much? Let’s look at GitHub

- Pytorch was the fastest growing software project
- Universities were 5 of the top 10 contributing organisations
- Literally billions of lines of code

96 M+ repositories

hosted on GitHub, 40% more than last year. Almost one third of all repositories were created in the last year.

Sharing is key to reproducibility

- Improves transparency
- Improves understanding
- Elimination of errors
- Encourages collaboration
- Easier on-ramping
- Improves trust

“Deep intellectual contributions now encoded only in software” – Stodden

“Scholarship is the full software environment, code and data, that produced the result” - Claerbout
Of 601 papers in ACM Computer Science journals and proceedings, only 85 provided a link to software. For 176 the software could not be obtained.

Collberg, Proebsting, Warren, University of Arizona TR 14-04, 2015
http://reproducibility.cs.arizona.edu/v2/RepeatabilityTR.pdf
In 2011 Science changed its editorial policies:

“We require that all computer code used for modeling and/or data analysis that is not commercially available be deposited in a publicly accessible repository upon publication.”

“After publication, all reasonable requests for data, code, or materials must be fulfilled.”

Stoddin, Seiler, Ma. An empirical analysis of journal policy effectiveness for computational reproducibility
https://doi.org/10.1073/pnas.1708290115
Table 1. Responses to emailed requests (n = 180)

<table>
<thead>
<tr>
<th>Type of response</th>
<th>Count</th>
<th>Percent, %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Did not share data or code:</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Contact another person</td>
<td>20</td>
<td>11</td>
</tr>
<tr>
<td>Asked for reasons</td>
<td>12</td>
<td>7</td>
</tr>
<tr>
<td>Refusal to share</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Directed back to supplement</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>Unfulfilled promise to follow up</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Impossible to share</td>
<td>65</td>
<td>36</td>
</tr>
<tr>
<td>Shared data and code</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Email bounced</td>
<td>46</td>
<td>26</td>
</tr>
</tbody>
</table>

When you approach a PI for the source codes and raw data, you better explain who you are, whom you work for, why you need the data and what you are going to do with it.

I have to say that this is a very unusual request without any explanation! Please ask your supervisor to send me an email with a detailed, and I mean detailed, explanation.

The data files remain our property and are not deposited for free access. Please, let me know the purpose you want to get the file and we will see how we can help you.

Normally we do not provide this kind of information to people we do not know. It might be that you want to check the data analysis, and that might be of some use to us, but only if you publish your findings while properly referring to us.

Thank you for your interest in our paper. For the [redacted] calculations I used my own code, and there is no public version of this code, which could be downloaded. Since this code is not very user-friendly and is under constant development I prefer not to share this code.

Stodden, Seiler, Ma. An empirical analysis of journal policy effectiveness for computational reproducibility

https://doi.org/10.1073/pnas.1708290115

Software Sustainability Institute
Preservation vs Sustainability

“IPK Gatersleben cold storage” by Dag Terje Filip Endresen

“1 Waverley Community Garden” by d-olwen-dee

Software Sustainability Institute
Purposes of preservation

- Achieve legal compliance
  - e.g. LOTAR (aerospace)

- Create heritage value
  - e.g. Salman Rushdie archive (literature)

- Enable continued access to data and services
  - e.g. aphid database (biosciences)

- Encourage software reuse
  - e.g. DOE Code (general)

Decisions, decisions

• There are several approaches that could be classed as software preservation
• The choice depends on a number of factors, which change through time

Software repository lifecycle

- **Code repository**
  - GitHub
  - Bitbucket
  - GitLab
  - Mercurial
  - Subversion

- **Digital repository**
  - Zenodo
  - Figshare
  - Dryad
  - DSpace
  - EPrints

- **Archival repository**
  - Archivematica
  - Fedora
  - Preservica
  - Software Heritage
  - Internet Archive

Emphasis on collaboration
Contents rapidly changing
*Individual / Project*

Emphasis on sharing
Contents slowly changing
*Community*

Emphasis on integrity
Contents rarely changing
*Organisation / Institution*
What about the R-word?

- Reputation?
  - Most software preservation is aimed at reducing risk in some way
- Reproducible?
  - Reproducibility adds extra burden on the developer
- Reusable?
  - How easy is it to reuse software after X years?
Why deposit software?

Personal motivation:
- Repeatability
- Reproducibility
- Reuse
- Delegate storage / distribution responsibility

Personal benefits:
- Save time/effort

Research-related motivation:
- Historical Record
- Store more complete research ecosystem
- Peer review
- Repeatability
- Reproducibility
- Reuse

Research-related benefits:
- Promote trust
- Improve quality
- Get credit
- Get collaborators

Community benefits:
- Promote trust
- Improve quality

Funder/publisher requirement

When to deposit software?

Ask yourself...

- Publishing results derived from your software?
- Sharing your software?
- Project coming to an end?
- Project going into a hiatus?

Deposit now

Concerned that others may publish results based on your software before you do?

Apply an embargo to publish later

Where to deposit software?

Avoid popular but problematic options:
- Laptop
- Desktop
- USB
- Website (personal, project, department)
- Repository hosting service (GitHub, BitBucket, GitLab,...)

Use to manage software under active development

Choose a digital repository:

Identify your options:
- Institutional
- Publisher/funder mandated/recommended
- Community recommended
- General e.g. Zenodo, Figshare, Software Heritage

Assess your options:
- Does it issue persistent identifiers?
- Is it mandated by a publisher or funder?
- Is it longevity acceptable?
- If there is a fee, is this a one-off payment and can you afford it?
- Can it accommodate the size of your deposit?
- Are its policies & service agreements acceptable?
- Is it free or is there a fee?
- Is it accredited or certified?

What to deposit?

Research software is changing
What else is going on?

- The Turing Way

- Software Heritage
  - [https://www.softwareheritage.org/](https://www.softwareheritage.org/)

- Software Preservation Network
  - [https://www.softwarepreservationnetwork.org/](https://www.softwarepreservationnetwork.org/)

- Software Citation Repository Best Practices Taskforce

- PyRDM
  - [https://github.com/pyrdm/pyrdm](https://github.com/pyrdm/pyrdm)

- CodeMeta
  - [https://codemeta.github.io/](https://codemeta.github.io/)

- WholeTale
  - [https://wholetale.org/](https://wholetale.org/)
Take home messages

• Software often has complex dependencies / interactions
  ▪ Sensitive to changes in its environment
  ▪ May require expert knowledge outside of the team

• Preserving source code has become easier – but binaries and services are still hard, even with new technology

• We need to understand why to decide how we preserve a piece of software

Slides: https://doi.org/10.6084/m9.figshare.8088290
Brings new meaning to the term “unboxing video”
Acknowledgements

The SSI team/alumni:  
- Aleksandra Nenadic  
- Aleksandra Pawlik  
- Alexander Hay  
- Arno Proeme  
- Carole Goble  
- Claire Wyatt  
- Clem Hadfield  
- Dave De Roure  
- Devasena Prasad  
- Giacomo Peru  
- Graeme Smith  
- Iain Emsley  
- James Graham  
- John Robinson  
- Les Carr  
- Malcolm Atkinson  
- Malcolm Illingworth  

- Mark Parsons  
- Mike Jackson  
- Olivier Philippe  
- Priyanka Singh  
- Raniere Silva  
- Rob Baxter  
- Robin Wilson  
- Shoaib Sufi  
- Simon Hettrick  
- Stephen Crouch  
- Tim Parkinson  
- Toni Collis  
- Plus the SSI Fellows and RSE community  

Scientific software:  
- Dan Katz  
- Heather Piowowar  
- James Howison  
- Jeff Carver  
- Jennifer Schopf  
- Kaitlin Thaney  
- Martin Fenner  
- Victoria Stodden  
- WSSSPE community  

Software/Data Carpentry  
- Greg Wilson  
- Jonah Duckles  
- Tracy Teal  
- Instructor Community  

Supported by the UK Research Councils through grants EP/H043160/1, EP/N006410/1 and EP/S021779/1.  
Additional project funding received from Jisc.
About the Institute
A national facility for cultivating better, more sustainable, research software to enable world-class research

• Software reaches boundaries in its development cycle that prevent improvement, growth and adoption
• Providing the expertise and services needed to negotiate to the next stage
• Developing the policy and tools to support the community developing and using research software

Supported by the UK Research Councils through grants EP/H043160/1, EP/N006410/1 and EP/S021779/1
Software Sustainability Institute

Software

Helping the community to develop software that meets the needs of reliable, reproducible, and reusable research

Training

Delivering essential software skills to researchers via CDTs, institutions & doctoral schools

Outreach

Exploiting our platform to enable engagement, delivery & uptake

Policy

Collecting evidence on the community’s software use & sharing with stakeholders

Community

Bringing together the right people to understand and address topical issues