www.software.ac.uk

# 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:*
BBSRC
Arts & Humanities Research Council
EPSRC Engineering and Physical Sciences Research Council
ESRC Economic & Social Research Council
MRC Medical Research Council
NERC Science of the Environment
Science & Technology Facilities Council

# Andy Warhol's Time Capsules

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

**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
**DATE RANGE**: 1977-March 1982, undated; Bulk: February-March 1982
**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

---

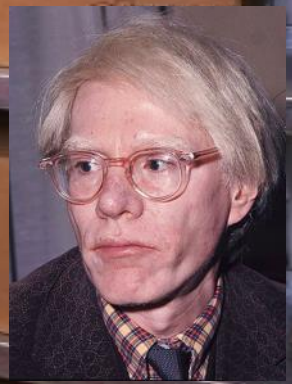
**ARRANGEMENT SUMMARY**

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/Textiles, 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.

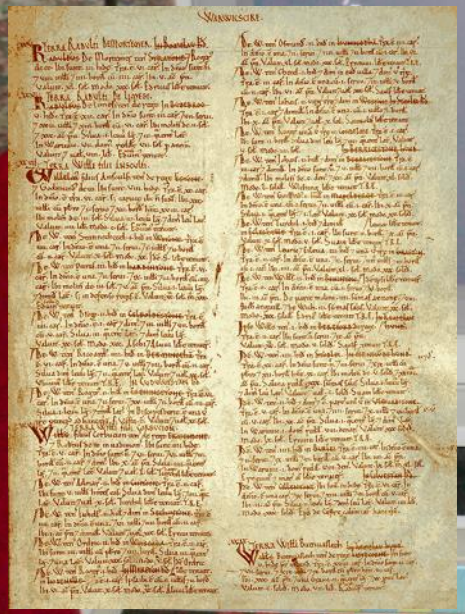
---

**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.



# Digital Domesday?



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

# DOMESDAY RELOADED



Picture of the day

## Examples of transport in 1985.

Mr Mart's horse drawn carriages and Mrs Margetts' Sinclair C5 are often seen in Long Bennington. Notice background building with mansard roof.

## 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.

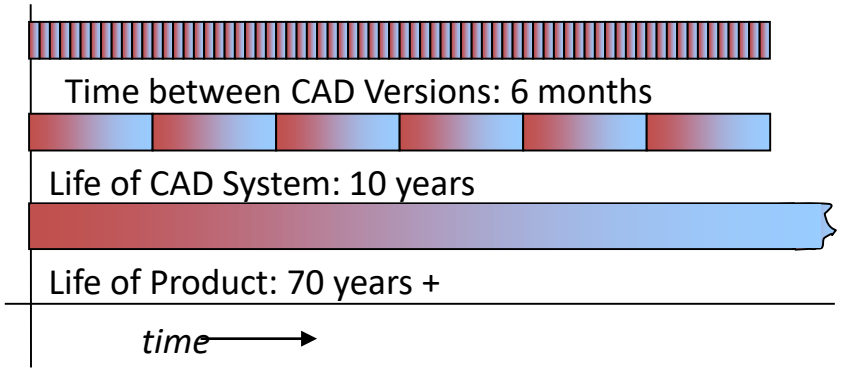
**Read more about the story of Domesday here**

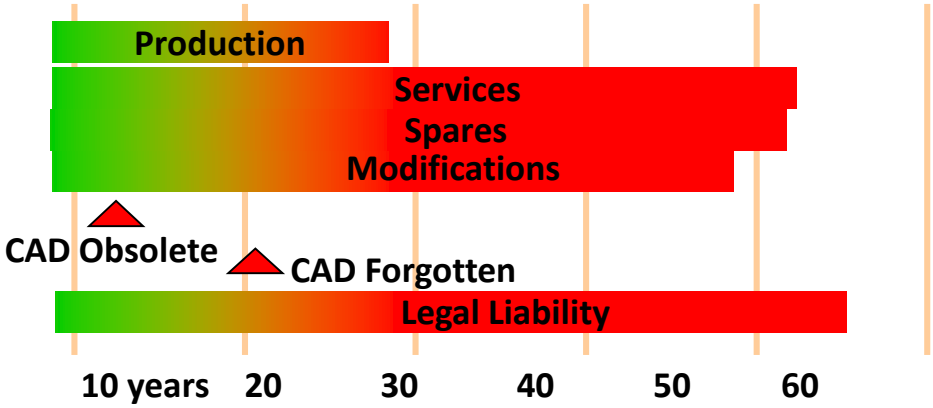
## Search the Domesday Site



Type place/postcode Place Content

# Archiving Models



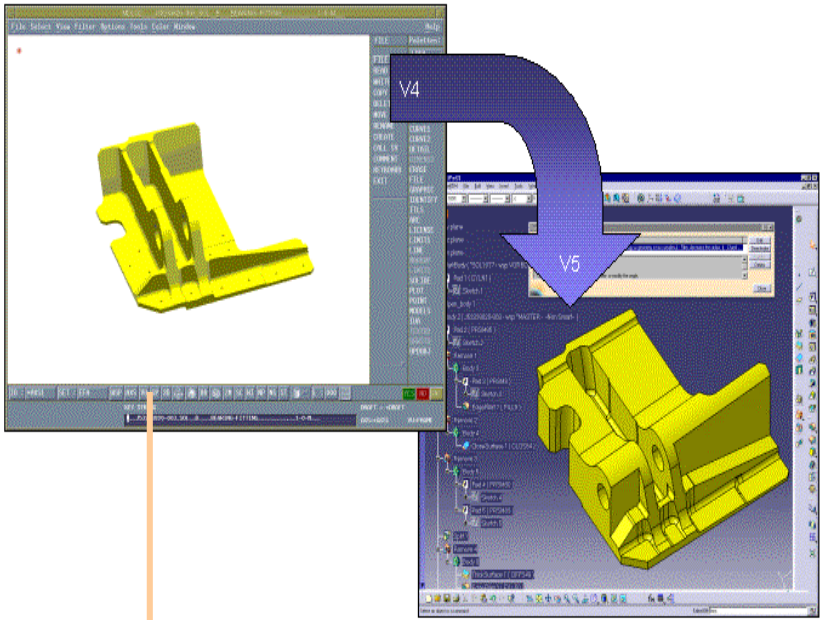
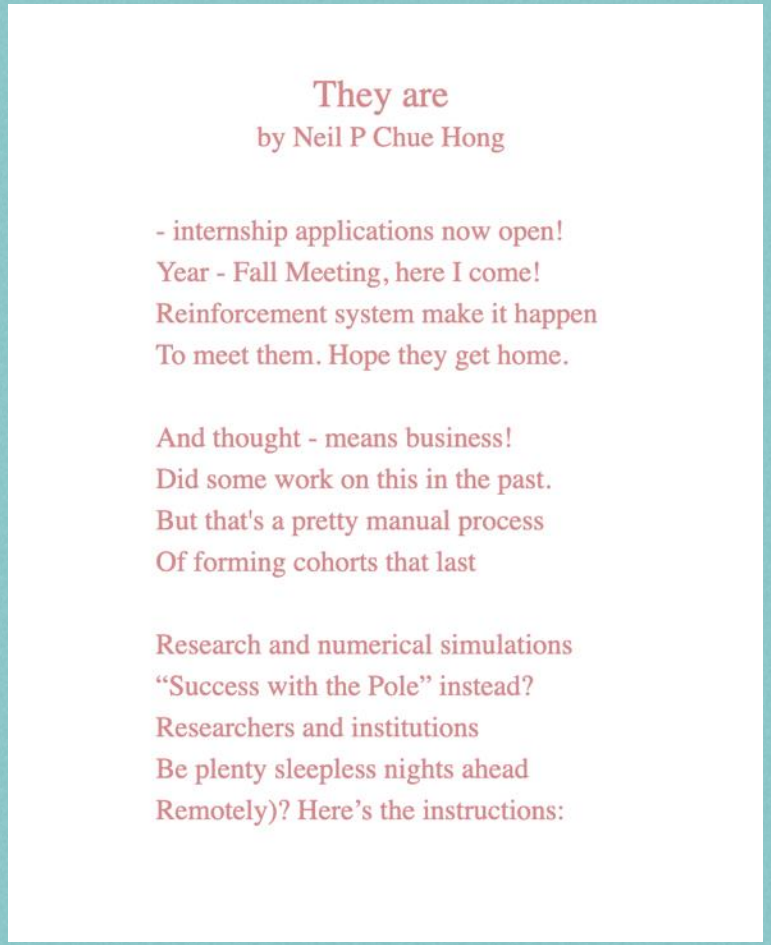




Image courtesy PDES Inc
Slide from Sean Barker, BAE SYSTEMS, DPC Designed to Last
https://slideplayer.com/slide/10521357/

# Interlinked Services

poetweet

@ npch



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

Re: Dark Star Inbox x enron x

**Richard B** <everyone@enron.email>
to me

After further thought, it seems to me that in light of our fear of litigation w/ 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.

Judge Scott Link Campaign Inbox x enron x

**Richard B** <everyone@enron.email>
to me

This is our judge in the Beeson case. I don't like him as a judge but I would recommend that we gave him $500-1000.
---------- Forwarded by Richard B Sanders/HOU/ECT on 07/09/99 05:07 PM ----------

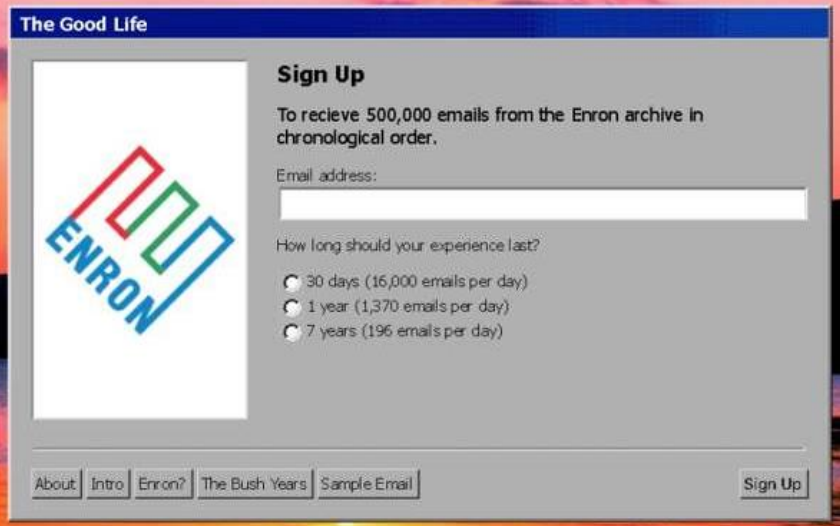
Enron Capital & Trade Resources Corp.

From: HmlangeTX@aol.com 07/09/99 03:46 PM

Good morning!

**Mark - ECT 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 16, 15, 14, 8, 7, 2 and 1 hit a little too close to home.
>
> TOP 22 SIGNS THAT YOU HAVE HAD TOO MUCH OF THE '90s
>
> 22. Cleaning up the dining area means getting the fast-food bags out of the back seat of your car.
>
> 21. Your reason for not staying in touch with family is that they do not have e-mail addresses.
>
> 20. Keeping up with sports entails adding ESPN's home page to your bookmarks.
>
> 19. You have a "to do" list that includes entries for lunch and bathroom breaks, and they are the ones that never get crossed off.
>
> 18. You have actually faxed your Christmas list to your parents.
>
> 17. Pick-up lines now include a reference to liquid assets and capital gains.
>
> 16. You consider second-day air delivery painfully slow.
>
> 15. You assume the question "to valet-park or not" is rhetorical.
>
> 14. You refer to your dining-room table as the flat filing cabinet.
>




WWF Inbox x enron x

**Terence H**
to me

I got the message that we are headed for Zurich for the WWF meeting in November. This is a very delicate situation in that WWF will use whatever you say, even the hint of the slightest committement, against Enron in the future. As you know, I have found this group the most dishonest, back dealing NGO I have ever dealt with.When Shell has these meetings, they have my counterpart meet ahead of time with the the WWF top staff person and work out an agenda. For example, you will have to be throughly briefed on the Bolivian situation before we go in. I also would find it counter productive to include any of the low level WWF staff that went to this meeting when Shell held it last. Remember, this is the group that publicly announced that Enron has gotten away with murder for years and we are going to get them. They also have te goal of wanting to lock us up in a world wide partnership. This is the last groupo I would do this with.

I guess this is a long way of saying that we need to go in their well prepared with our eyes wide open. Did you commit to any agenda when you set up the meeting? I will contact their staff and see what they want to discuss and when I get back to the states, I start pulling materials together.

Derivatives Inbox x enron x

**Gerald** <everyone@enron.email>
to me

Lisa, It is extremely hard to watch you struggle through this. It is hard for me also. I sit up here trying to work when the reality of our current situation weighs on me like a ton of bricks. It is hard for me not to break down a couple of times a day, but the pace of my work usually forces me through those times.

I don't know if my calling you, asking you to lunch, or holding your hand, helps or hurts at this time. But I am not going to let us give up on each other. My feelings for you run too deep to just try and bury them so that I don't have to think about all the pain.

We are both struggling to find a place to start. I want to start by just letting you know that not a day goes by without me thinking of a time when we can put this behind us. I am not sure how to get there right now either. However, during those few precious time lately when I get to look into your eyes, I see the true Lisa (changed as she is by the last couple of years), but still Lisa nonetheless. I need to get to know this changed true Lisa, whatever happens in the near future.

I felt I just to had to write this email.

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

jupyter nbviewer

JUPYTER FAQ

# Transparent Journalism

## 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 critique 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 pundits 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 demarcating 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 use "Budget" on both sides of the equation, which is a big no-no. Remember, we constructed ROI as $(Revenue - Budget)/Budget$ and Profit as $Revenue - Budget$ 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 criteria and Revenue. Better roles for women translates into better revenue, even controlling for the fact that bigger budgets also create more revenue. This model also explains approximately 24% of the variance, versus 15% in the article's model suggesting it's doing a better job modeling the relationship of Bechdel test and financial performance.

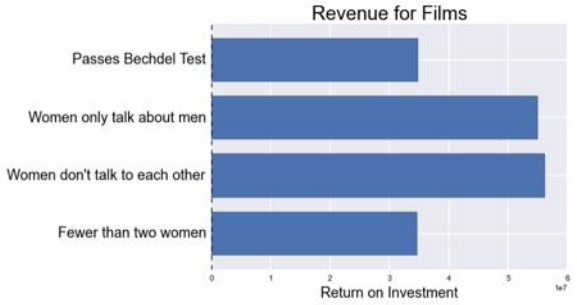
n [48]:
```
# Make the bar-plot
df['Adj_Revenue'].groupby(df['rating']).agg(np.median).plot(kind='barh')

plt.yticks(plt.yticks()[0],["Fewer than two women",
                            "Women don't talk to each other",
                            'Women only talk about men',
                            'Passes Bechdel Test'
                            ],fontsize=18)
plt.xlabel('Return on Investment',fontsize=18)
plt.title('Revenue for Films',fontsize=24)

plt.ylabel('')

# Estimate the model
ed_m3 = smf.ols(formula='log(Adj_Revenue+1) ~ C(rating)  + log(Adj_Budget+1)', data=df).fit()
print ed_m3.summary()
```

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | log(Adj_Revenue + 1) | R-squared: | 0.244 |
| Model: | OLS | Adj. R-squared: | 0.242 |
| Method: | Least Squares | F-statistic: | 127.2 |
| Date: | Mon, 07 Apr 2014 | Prob (F-statistic): | 3.12e-94 |
| Time: | 09:19:20 | Log-Likelihood: | -3338.6 |
| No. Observations: | 1583 | AIC: | 6687. |
| Df Residuals: | 1578 | BIC: | 6714. |
| Df Model: | 4 | | |

| | coef | std err | t | P>\|t\| | [95.0% Conf. Int.] | |
|---|---|---|---|---|---|---|
| Intercept | 1.8864 | 0.709 | 2.661 | 0.008 | 0.496 | 3.277 |
| C(rating)[T.1.0] | 0.2741 | 0.203 | 1.351 | 0.177 | -0.124 | 0.672 |
| C(rating)[T.2.0] | 0.6096 | 0.232 | 2.626 | 0.009 | 0.154 | 1.065 |
| C(rating)[T.3.0] | 0.4483 | 0.193 | 2.318 | 0.021 | 0.069 | 0.828 |
| log(Adj_Budget + 1) | 0.8825 | 0.039 | 22.438 | 0.000 | 0.805 | 0.960 |

| | | | |
|---|---|---|---|
| Omnibus: | 1650.678 | Durbin-Watson: | 1.998 |
| Prob(Omnibus): | 0.000 | Jarque-Bera (JB): | 111167.694 |
| Skew: | -5.042 | Prob(JB): | 0.00 |
| Kurtosis: | 42.796 | Cond. No. | 248. |




Brian C. Keegan "The Need for Openness in Data Journalism"
https://www.brianckeegan.com/2014/04/the-need-for-openness-in-data-journalism/
https://nbviewer.jupyter.org/github/brianckeegan/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


https://octoverse.github.com/ (2018)

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

# Sharing is key to reproducibility

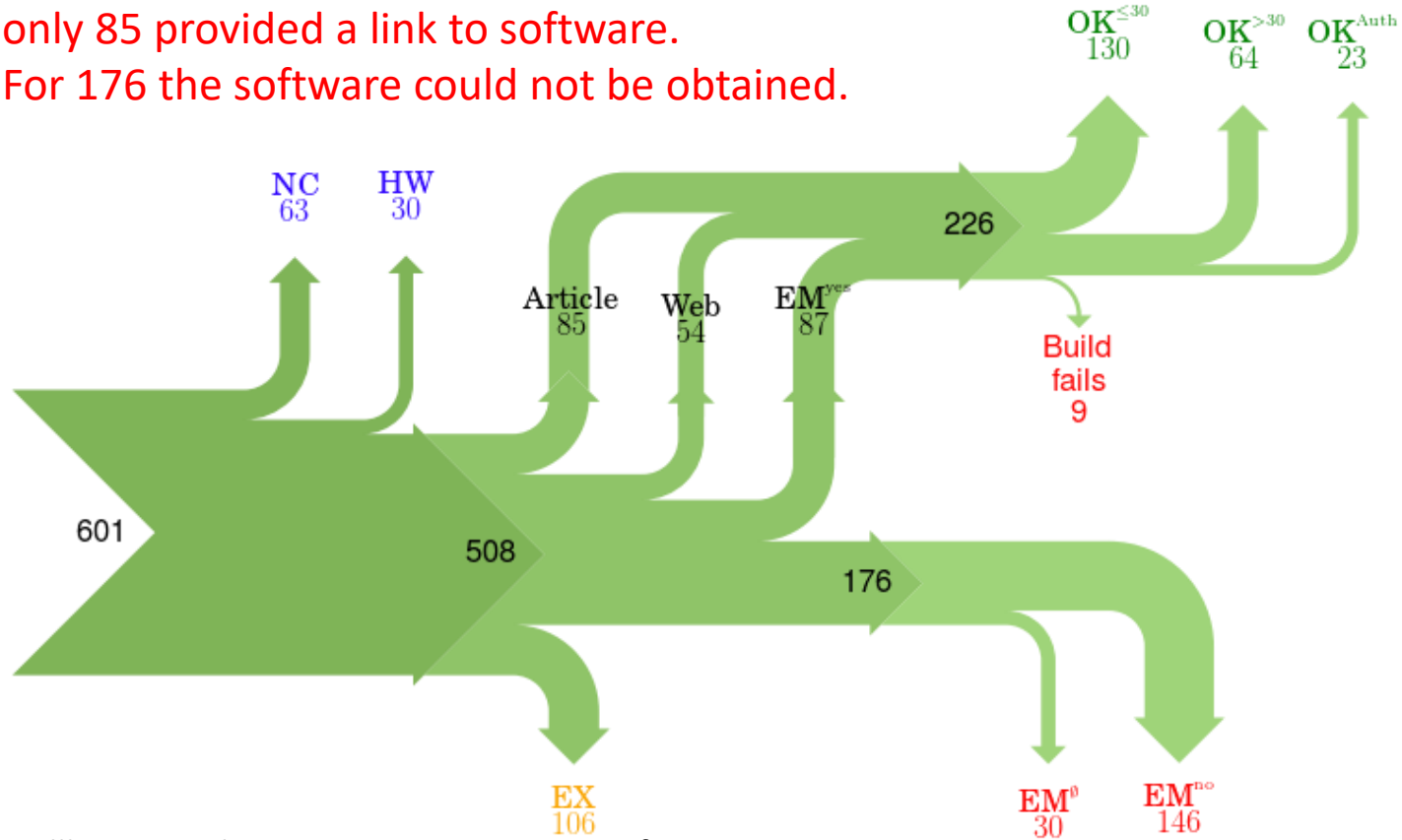

- Improves transparency
- Improves understanding
- Elimination of errors
- Encourages collaboration
- Easier on-ramping

- 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

www.software.ac.uk

# Culture change is hard


Science
AAAS

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."*

**Stodden, Seiler, Ma. An empirical analysis of journal policy effectiveness for computational reproducibility**
https://doi.org/10.1073/pnas.1708290115

# Culture change is hard

www.software.ac.uk

Science
AAAS

**Table 1. Responses to emailed requests ($n$ = 180)**

| Type of response | Count | Percent, % |
|---|---|---|
| Did not share data or code: | | |
| Contact another person | 20 | 11 |
| Asked for reasons | 20 | 11 |
| Refusal to share | 12 | 7 |
| Directed back to supplement | 6 | 3 |
| Unfulfilled promise to follow up | 5 | 3 |
| Impossible to share | 3 | 2 |
| Shared data and code | 65 | 36 |
| Email bounced | 3 | 2 |
| No response | 46 | 26 |

**Stodden, Seiler, Ma. An empirical analysis of journal policy effectiveness for computational reproducibility**
https://doi.org/10.1073/pnas.1708290115

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 remains 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.

# Preservation vs Sustainability

www.software.ac.uk

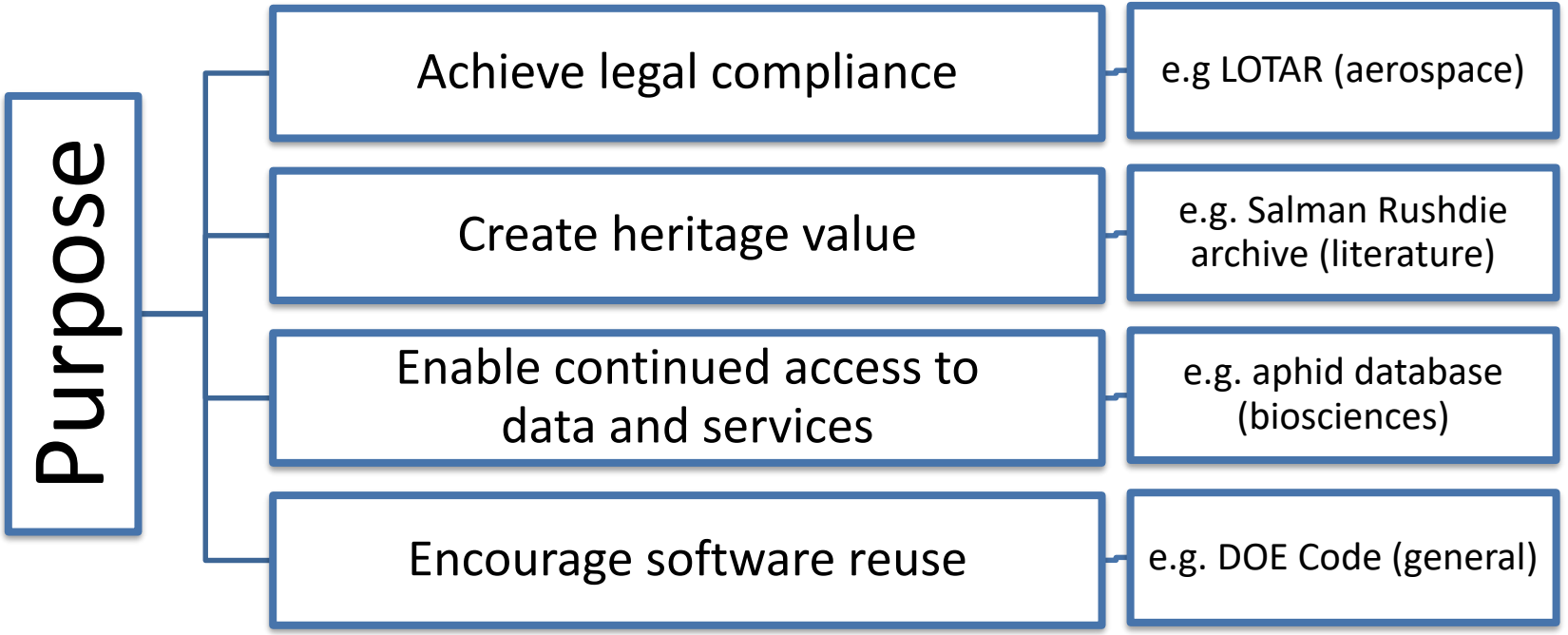
"IPK Gatersleben cold storage" by Dag Terje Filip Endresen

"1 Waverley Community Garden" by d-olwen-dee

# Purposes of preservation

www.software.ac.uk

Jisc




Chue Hong et al. (2010). Software Preservation Benefits Framework (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.2561972

Software Sustainability Institute

# Decisions, decisions

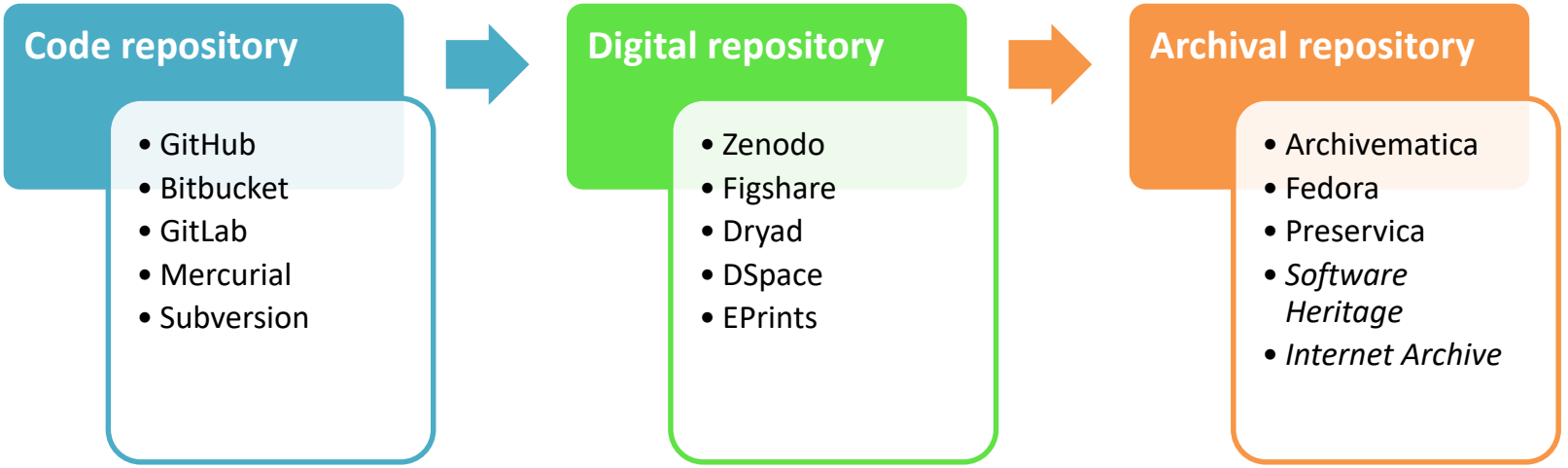
www.software.ac.uk

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


Chue Hong et al. (2010). Software Preservation Benefits Framework (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.2561972

# Software repository lifecycle

www.software.ac.uk




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?

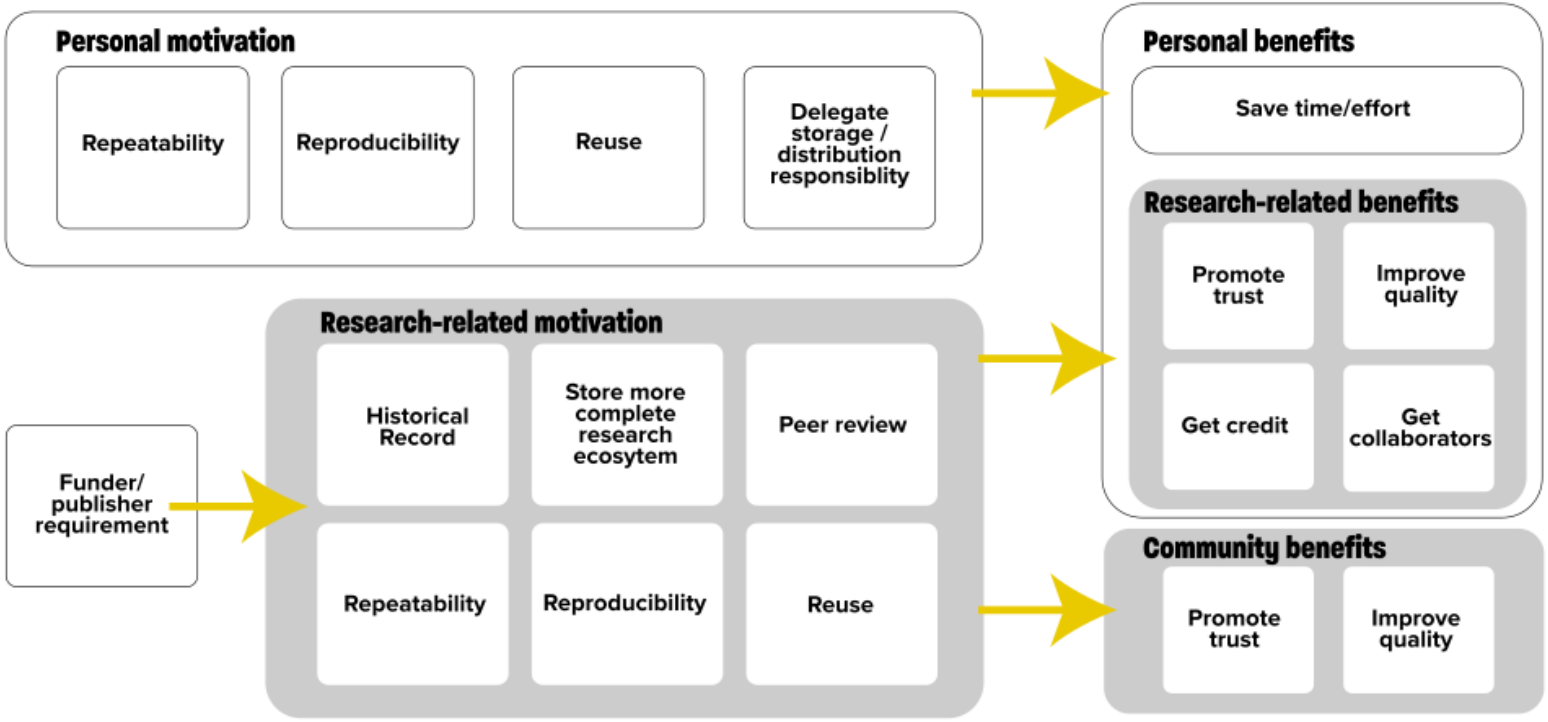

- Reputation?
  - Most software preservation is aimed at reducing risk in some way
- Reproducible?
  - Reproducibility adds extra burden on the developer
- Reusable?
  - How easy is it to reuse software after X years?



# Why deposit software?

www.software.ac.uk

Jisc

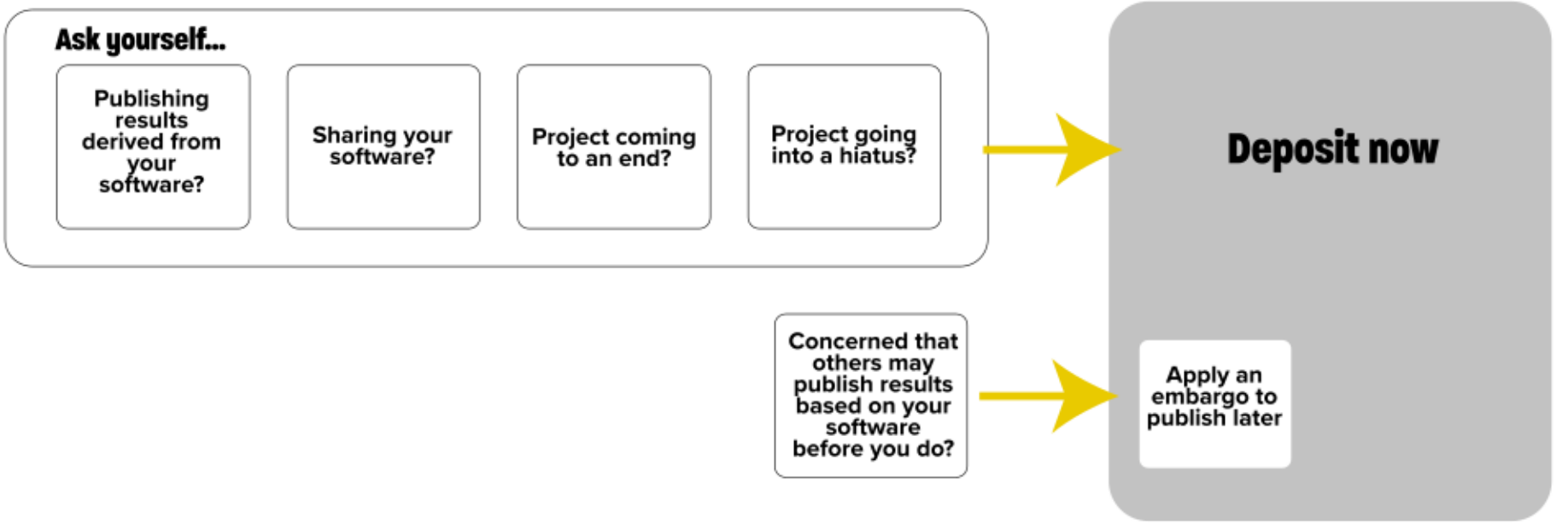



Michael Jackson. (2018). Software Deposit: Why deposit software (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1327333 Software Sustainability Institute

# When to deposit software?

www.software.ac.uk

Jisc

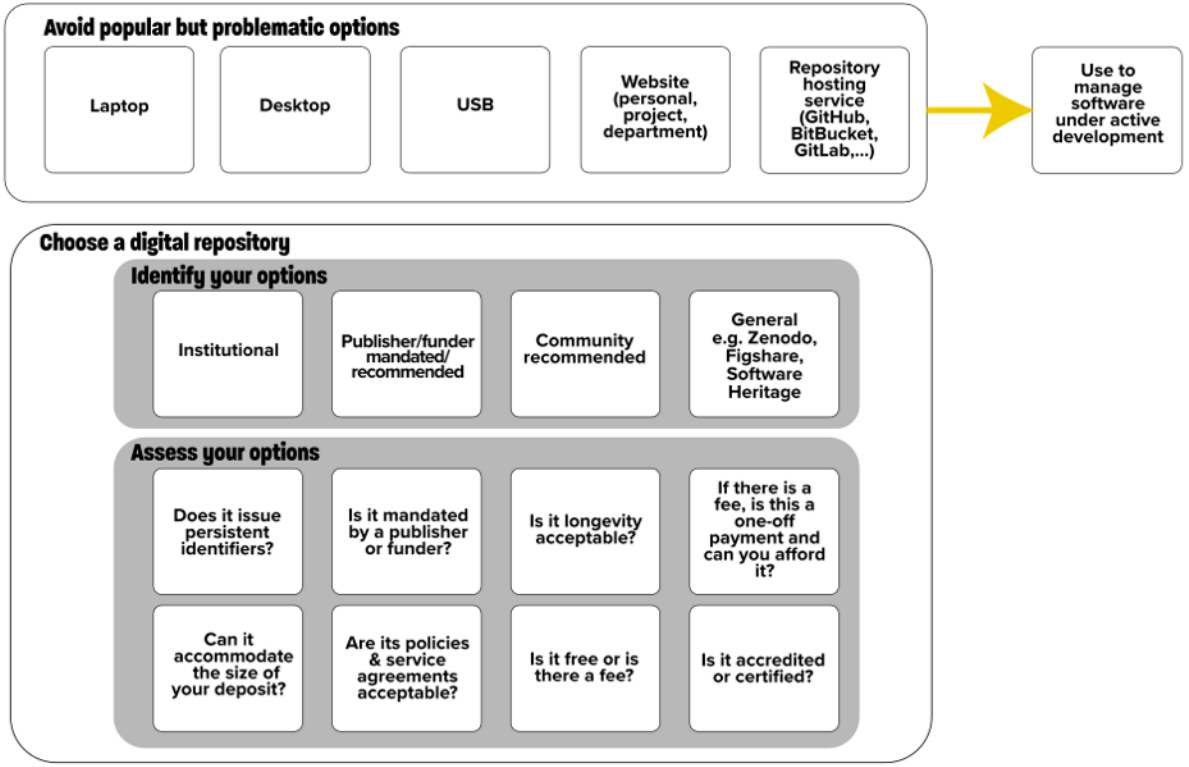



Michael Jackson. (2018). Software Deposit: When to deposit software (Version 1.0).
Zenodo. http://doi.org/10.5281/zenodo.1327331 Software Sustainability Institute

# Where to deposit software?

www.software.ac.uk

Jisc

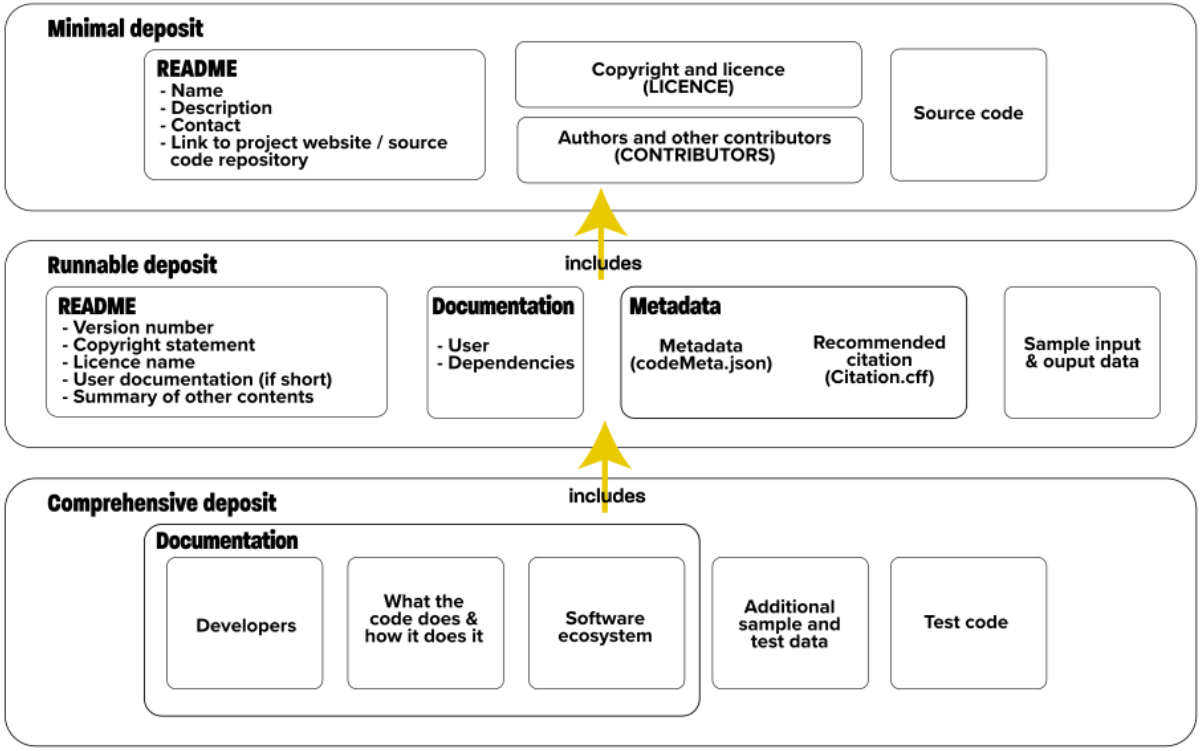



Michael Jackson. (2018). Software Deposit: Where to deposit software (Version 1.0). Zenodo. http://doi.org/10.5281/zenodo.1327329 Software Sustainability Institute

# What to deposit?

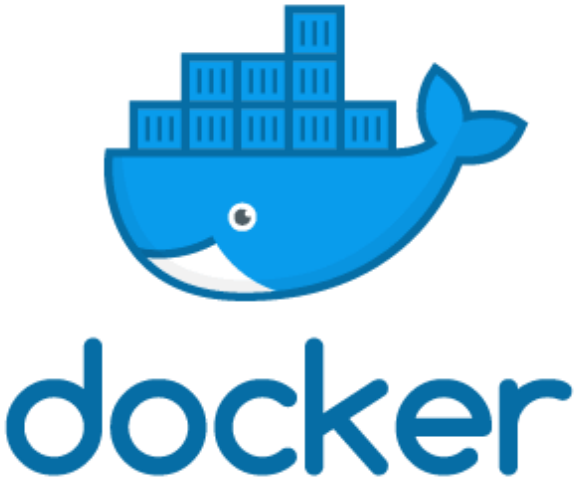
www.software.ac.uk

Jisc

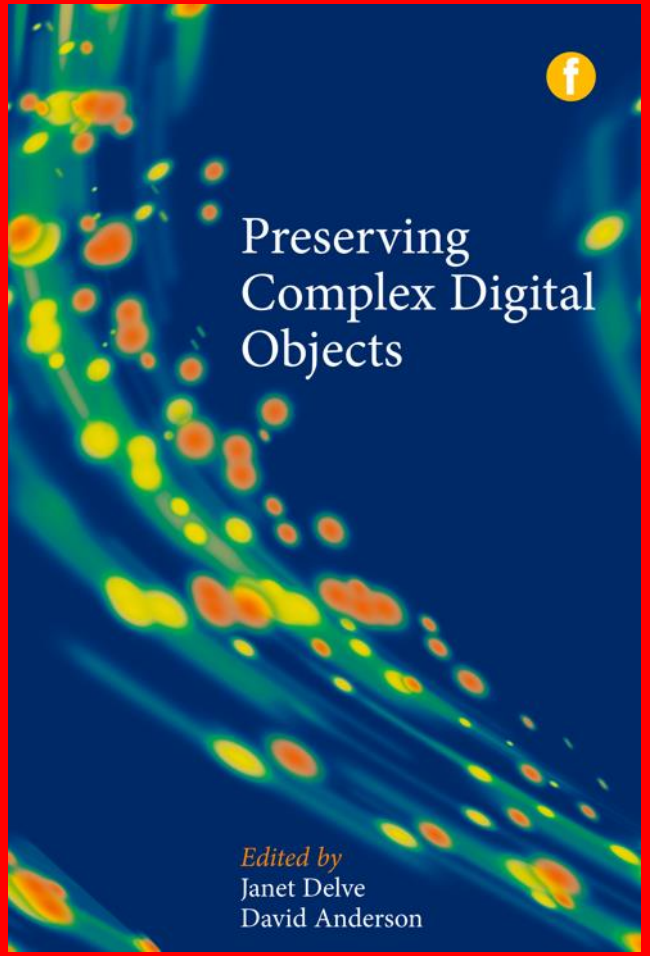



Michael Jackson. (2018). Software Deposit: What to deposit (Version 1.0).
Zenodo. http://doi.org/10.5281/zenodo.1327325 Software Sustainability Institute

# Research software is changing








Software
**Preservation**
Network

www.software.ac.uk

Software Heritage
THE GREAT LIBRARY OF SOURCE CODE

DigitalPreservationCoalition

WHOLE
TALE

# What else is going on?


- The Turing Way
  - https://www.turing.ac.uk/research/research-projects/turing-way-handbook-reproducible-data-science
- Software Heritage
  - https://www.softwareheritage.org/
- Software Preservation Network
  - https://www.softwarepreservationnetwork.org/
- Software Citation Repository Best Practices Taskforce
  - https://github.com/force11/force11-sciwg/issues/59
- PyRDM
  - https://github.com/pyrdm/pyrdm
- CodeMeta
  - https://codemeta.github.io/
- WholeTale
  - https://wholetale.org/

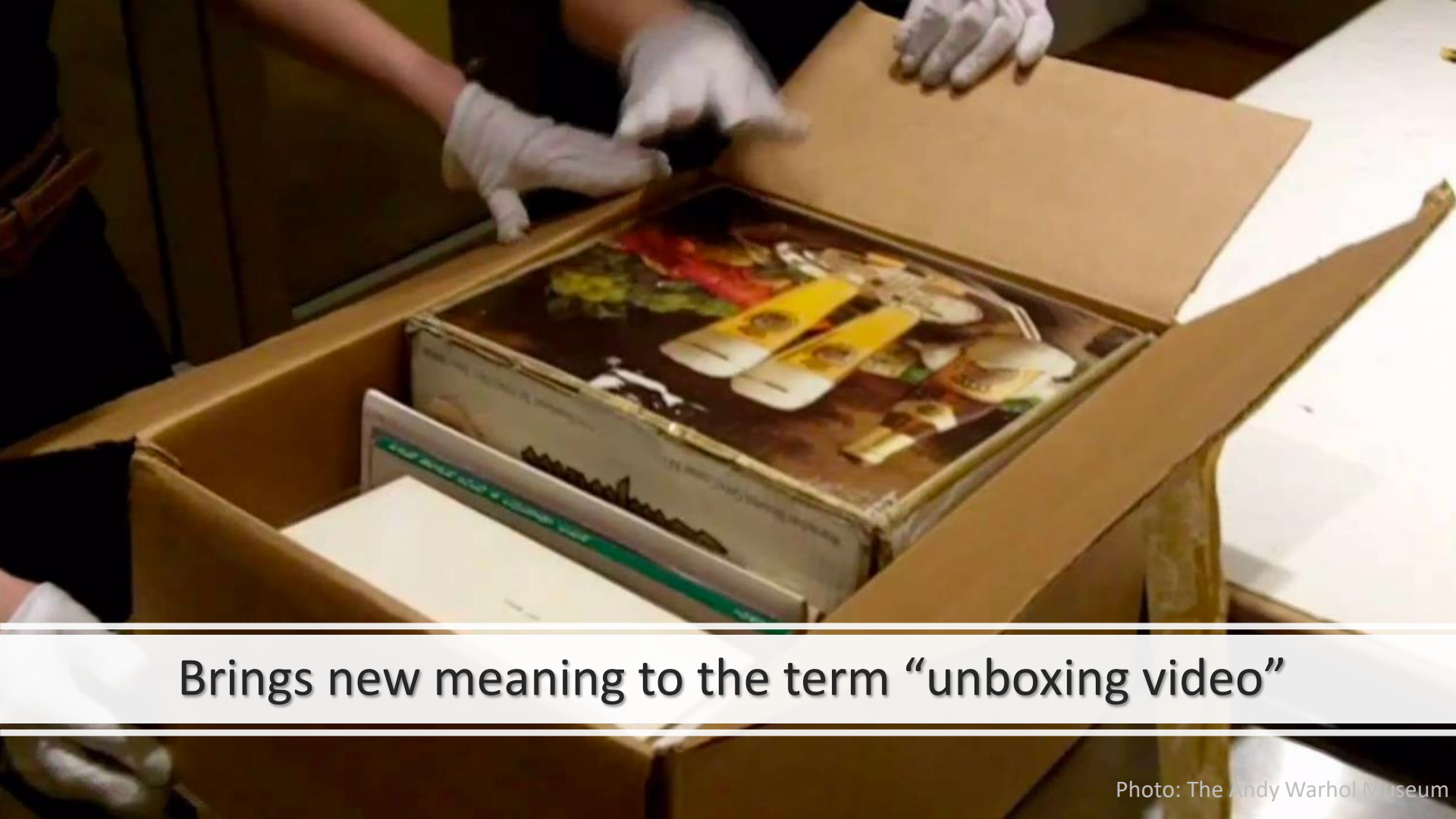


# Take home messages


- 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”

Photo: The Andy Warhol Museum

# 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*
- Mario Antonioletti
- 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.


Software Sustainability Institute

www.software.ac.uk


# About the Institute

# Software Sustainability Institute

www.software.ac.uk

## 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


Arts & Humanities Research Council

BBSRC

EPSRC
Engineering and Physical Sciences Research Council

E·S·R·C Economic & Social Research Council

MRC Medical Research Council

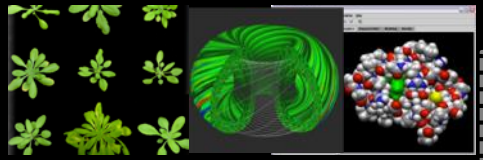
NERC Science of the Environment

Science & Technology Facilities Council

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

## Software

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



## Training

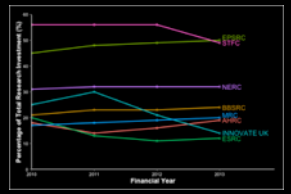
Delivering essential software skills to researchers via CDTs, institutions & doctoral schools

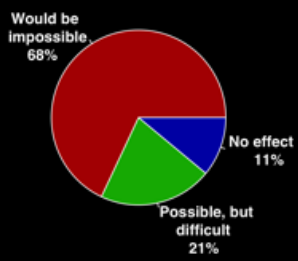

## Outreach

Exploiting our platform to enable engagement, delivery & uptake

## Policy

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

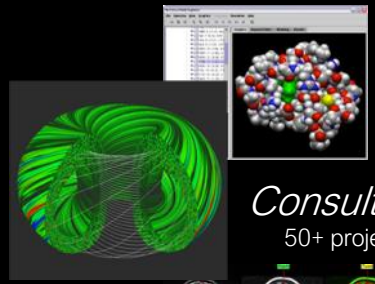



## Community

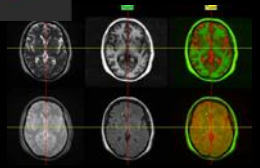
Bringing together the right people to understand and address address topical issues

# Software



## Consultancy

50+ projects



# Advice

130+ evaluations
4 surgeries

# Training


## Courses

35+ UK SWC workshops
1000+ learners

## Guides

80+ guides
50,000 readers

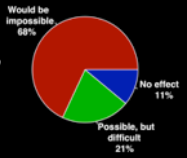
# Outreach

## Website & blog

150+ contributed articles
20,000 unique visitors per month
3,000 Twitter followers

## Research

740 researchers
50,000 grants analysed




## Campaigns


300+ RSEs engaged

BETTER SOFTWARE BETTER RESEARCH

2100 signatures

13 issues highlighted

Journal of open research software

## Workshops

20+ workshops organised

## Fellowship

61 domain ambassadors

# Policy

# Community