



STANFORD UNIVERSITY LIBRARIES

Measure All the (Web Archiving) Things!


Nicholas Taylor
[Web Archiving](#) Service Manager
[Stanford University Libraries](#)

[Archive-It Partner Meeting](#)
August 18, 2015

how many more websites are we archiving?



crawl report list


ARCHIVE-IT

[Collections](#)
[Crawls](#)
[Archives](#)

[Help](#)
[Submit Support Ticket](#)
Welcome, Nicholas Taylor

Home / Crawls

Stanford Humanities Lab

Select a crawl to compare.

[Crawl Reports](#)
[Current Crawls](#)
[Test Crawls](#)
[Scheduled Crawls](#)

Crawl Report List (1 to 100 of 199 Crawl Reports)

Type to Filter Crawl Reports


[Download Crawl Report List](#)
< 1 2 >


Archive-It: "Crawls for Account #198"

seeds for individual crawl



download seed list


Collections
Crawls
Archives
Help
Submit Support Ticket
Welcome, Nicholas Taylor

[Home](#) / [Crawls](#) / [99435](#) / [Seed Report](#)

Virtual Worlds and MMOS

Quarterly Crawl: 99435 | **Started:** February 9, 2014 6:41 PM | **Completed:** February 12, 2014 6:42 PM
 Show All Data
Show Only New Data

Crawl Overview
Seeds
Hosts
File Types


Total Data
60.7 GB
No Limit


New Data
25.6 GB


Total Docs
1,097,570
No Limit


New Docs
721,491


Unchanged Docs
376,079



60.7 GB
Total Capture

Data by Seed



Top 10 Seeds by Document Count

Seed	Document Count
http://warcraftmovies.com/	315,375
http://www.massively.com/	221,573
http://massively.tstiq.com/	139,466
http://www.massively.com/	120,766
http://www.massively.com/	102,234
http://www.massively.com/	66,464
http://www.massively.com/	47,224
http://www.massively.com/	44,763
http://www.massively.com/	39,152
http://www.massively.com/	261

Seed List (16 Seeds)

Download Seed List

Seed Url	Seed Status	Docs	Data	Wayback Link
http://warcraftmovies.com/	Crawled	139,466	20.8 GB	Wayback
http://www.massively.com/	Redirected	315,375	16.2 GB	Wayback
http://massively.tstiq.com/	Crawled	102,234	10.2 GB	Wayback

Archive-It: "[Seeds for Crawl #99435](#)"

downloaded seed list

seed-list.csv - Excel

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

A1 : X ✓ fx Seed Url

	A	B	C	D	E	F
1	Seed Url	Seed Status	Docs	Data	Wayback Link	
2	http://warcraftmovies.com/	Crawled	139466	22380303122	Wayback	
3	http://www.massively.com/	Redirected	315375	17445758365	Wayback	
4	http://elitistjerks.com/	Redirected	221573	6330914248	Wayback	
5	http://nwn.blogs.com/	Crawled	120766	5328314595	Wayback	
6	http://www.raphkoster.com/	Crawled	102234	3298974775	Wayback	
7	http://www.secondlifeinsider.com/	Redirected	66464	3298015087	Wayback	
8	http://terranova.blogs.com/	Crawled	44763	2882498436	Wayback	
9	http://secondlife.com/	Crawled	39152	2185351196	Wayback	
10	http://www.secondlifeherald.com/	Redirected	47224	1994286553	Wayback	
11	http://blue.cardplace.com/archive/	Crawled	261	4884176	Wayback	
12	http://www.worldofwarcraft.com/	Redirected	119	2660267	Wayback	
13	http://www.slobserver.com/	Crawled	67	1851605	Wayback	
14	http://www.metaversemessenger.com/	Crawled	54	611625	Wayback	
15	http://www.guildcafe.com/	Crawled	36	124715	Wayback	
16	http://www.secondseeker.com/	Crawled	14	40266	Wayback	
17	http://www.killerguides.com/	Not crawled (blocked by robots)	2	133	Wayback	
18						

seed-list

READY 100%

whew, that was easy!



oh, wait a minute...

seed lists are per crawl

well, how many crawls are there?

- 6 accounts
- oldest active since 2007
- 30+ collections
- **hundreds of crawls**

count and average not enough

- seeds move in and out of crawls
- seeds have different frequencies
- new seeds w/ new URLs for old seeds
- “university website” is many seeds

plus

- non Archive-It web archiving activity



“[Dichotomic Maples](#)” by [francoismi](#) under [CC BY-NC-SA 2.0](#)

“what gets measured, gets managed”



“Gudaauri still life” by [Carsten ten Brink](#) under [CC BY-NC-ND 2.0](#)

why measure?

- advocacy/outreach
- service modeling
- program assessment
- policy making
- staffing assessment
- grant support
- prioritization
- risk assessment

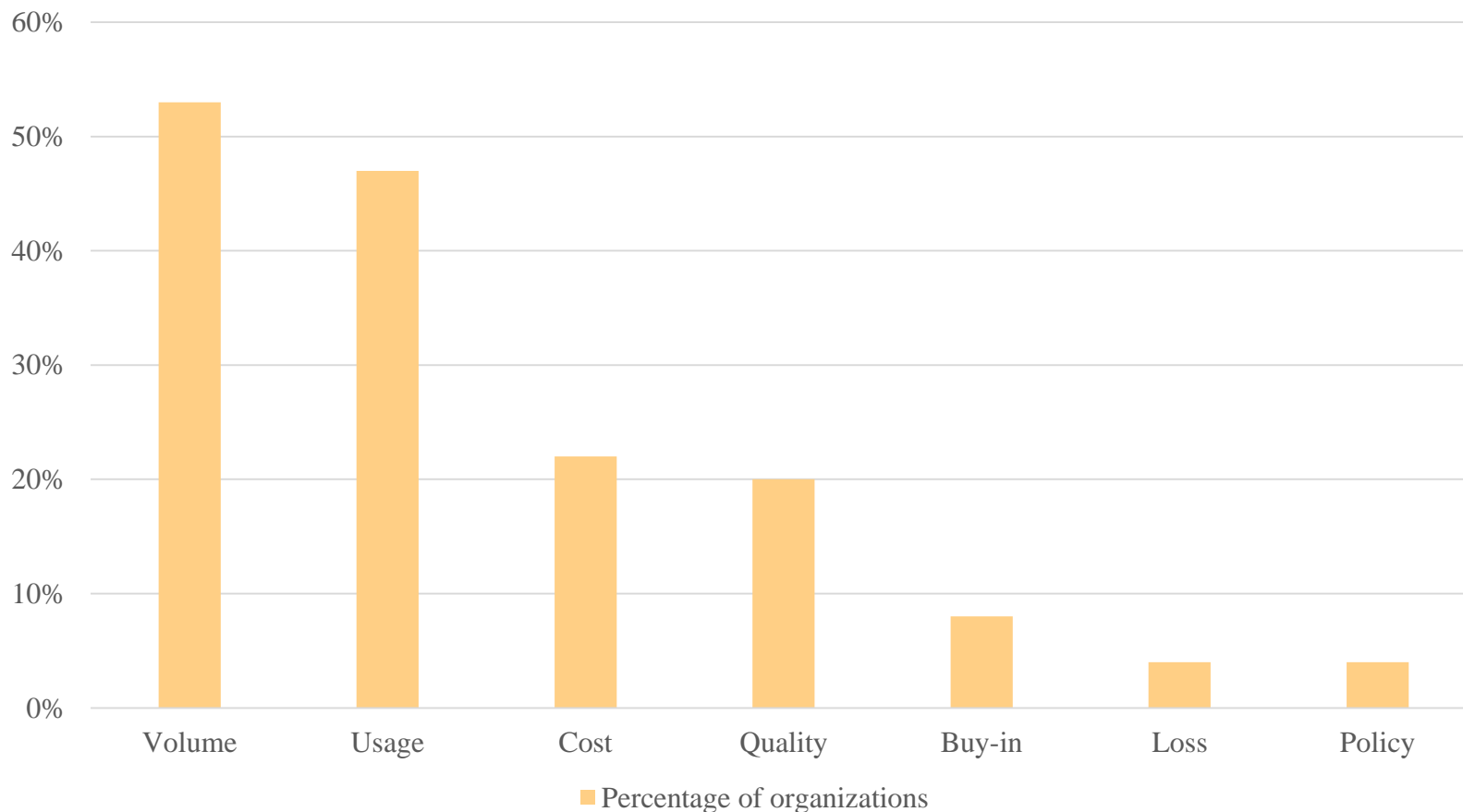


“[Measuring river depth](#)” by [epeirogenic](#) under [CC BY-NC 2.0](#)

what to measure?

- How to handle the data **volume**?
- What is the **usage** of web archives?
- How much does web archiving **cost**?
- How to assure the **quality** of archived content?
- How to secure institutional **buy-in**?
- How much **loss** have resources suffered?
- What is the impact of **policy** requirements?

community-valued metrics



[NDSA: "Web Archiving in the United States: a 2013 Survey"](#)

volume

- websites
 - captured
 - preserved
 - described
- data
 - captured
 - preserved
- objects
 - captured
 - preserved



“typography jumble” by [Bill Dickinson](#) under [CC BY-NC 2.0](#)

usage

- web analytics
 - visitors
 - visits
 - referers
- actual use cases
(who + how many?)
 - research
 - teaching
 - institutional legacy
 - compliance



“113/365 Days: A page from my heart” by [LaughingRhoda](#) under [CC BY-NC-ND 2.0](#)

cost

- external
 - out-payments for web archiving services
 - quota utilization
- internal
 - staff time, by activity
 - storage



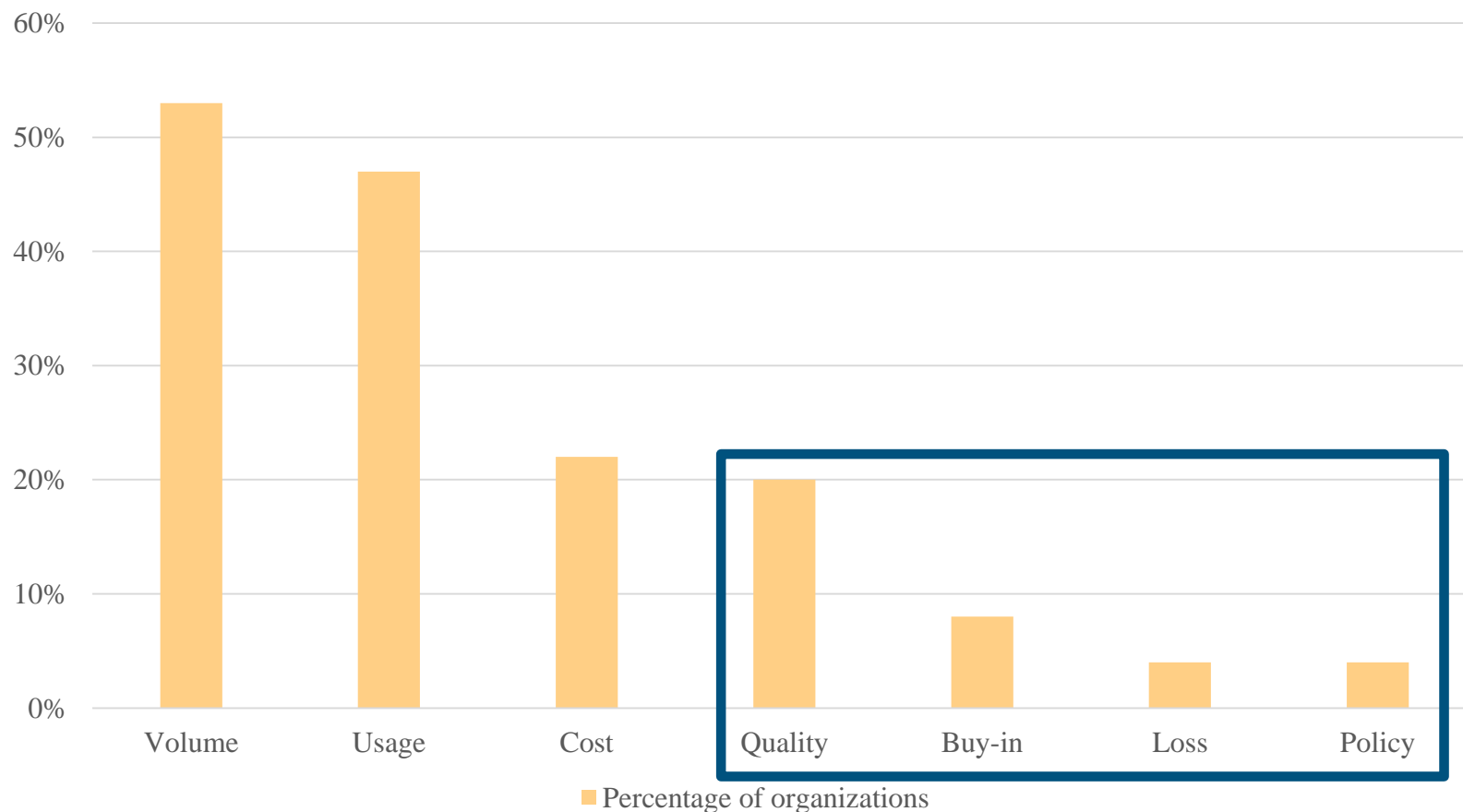
“[Largest square from a dollar bill](#)” by [origami madness](#) under [CC BY-NC 2.0](#)

performance

- **accessioning
throughput**
- **service request
turnaround**
- **collections/websites
w/ discovery records**
- time to regenerate
full-text index



community-valued...metrics?



[NDSA: "Web Archiving in the United States: a 2013 Survey"](#)

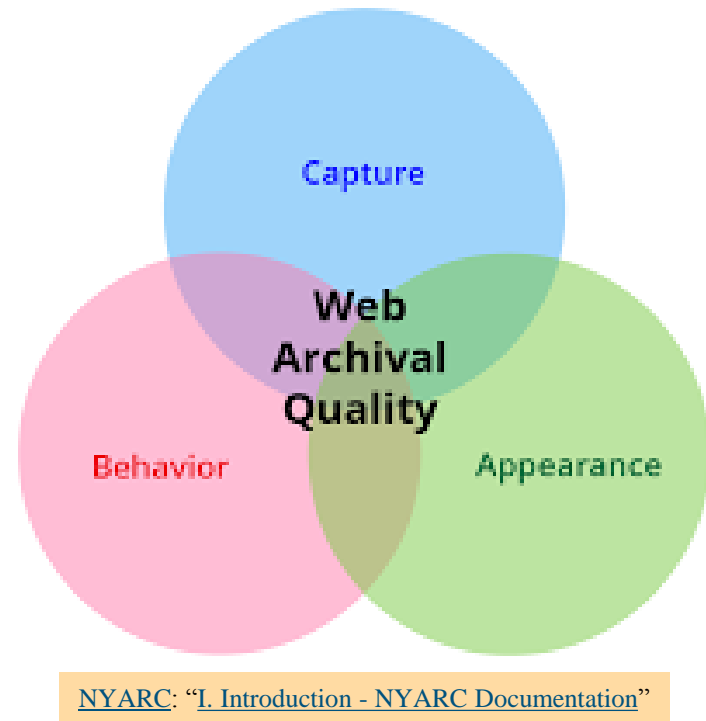
“not everything that counts can be counted”



“Ten Floods, Twenty-Five Trees, Nineteen Bubbles...” by [Flood G.](#) under [CC BY-NC-ND 2.0](#)

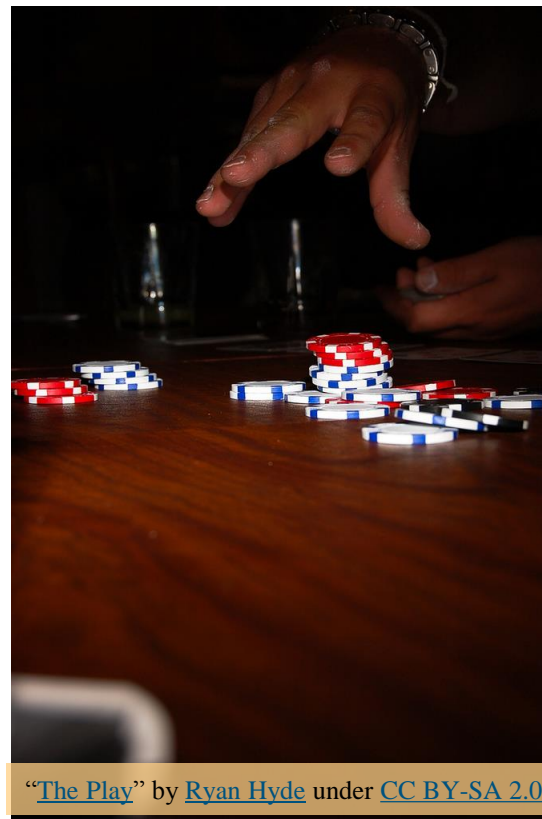
quality

- use case-specific?
- benchmark to ideal or to limits of tools?
- quantifiable metrics?
- existing metrics as proxies for quality?
- sampling approach?
- not just missing content but also collected junk



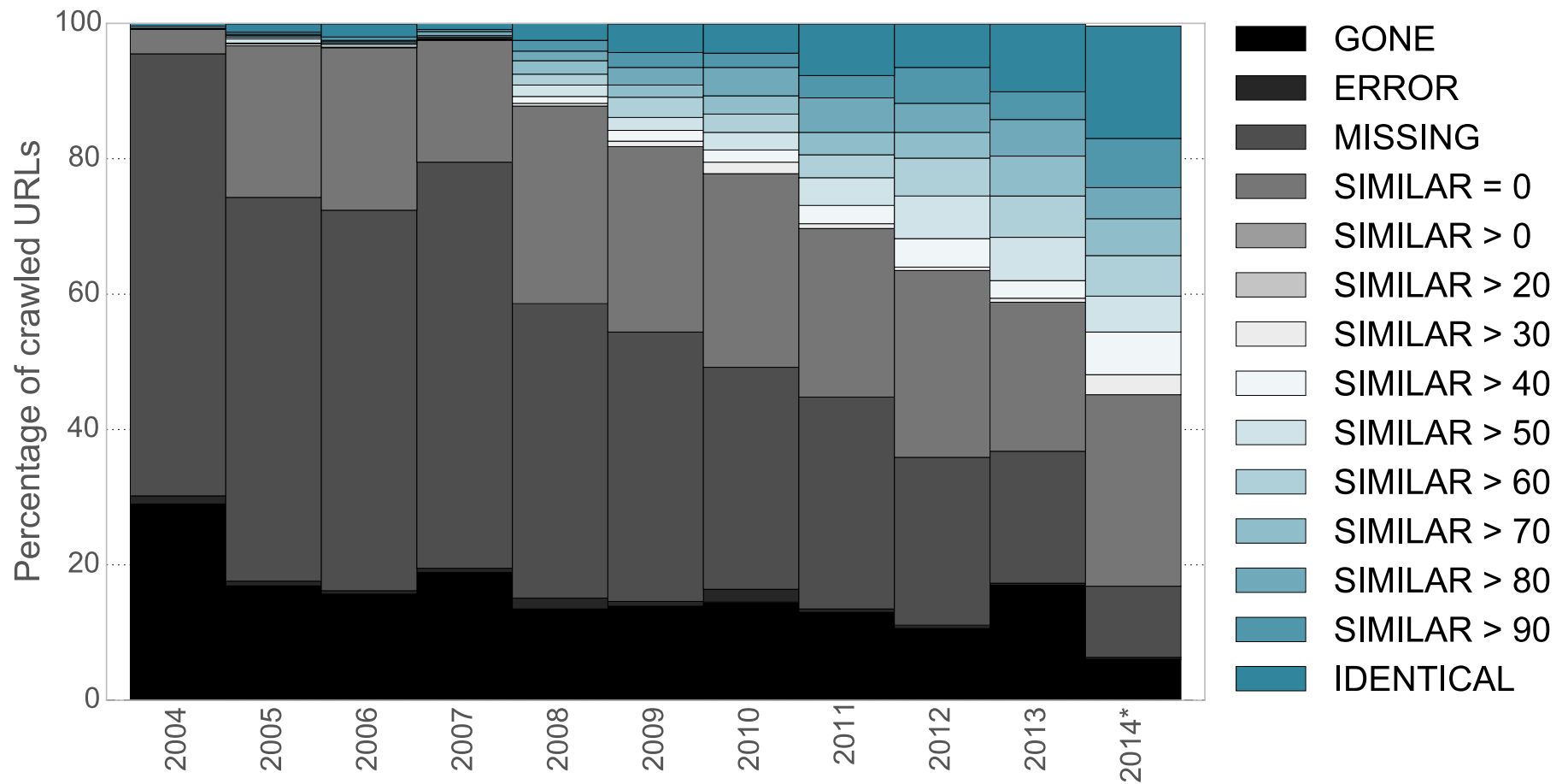
buy-in

- unique nominators?
- projects w/ web archiving component?
- budgetary commitments?
- resource commitments?
- charge for service?
- testimonials?



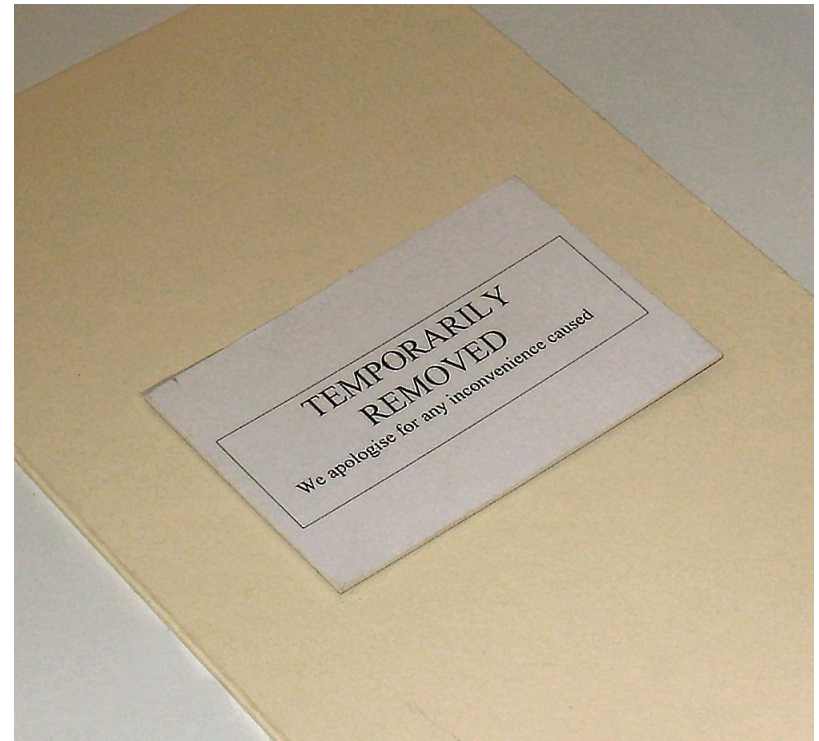
“The Play” by [Ryan Hyde](#) under [CC BY-SA 2.0](#)

loss



policy

- first capture under embargo
- opt-out requests
- takedown requests
- external environment



[“We apologise for any convenience - Update”](#) by [Alan Stanton](#) under [CC BY-SA 2.0](#)

better measures, measuring better



“Line Art Project #2 VIS3 UCSD” by [Mandy Jouan](#) under [CC BY-NC-ND 2.0](#)