HOW CLOSELY SHOULD WE WATCH THE BITS?

A long term media experiment protocol

Tyler Cline
J. Murrey Atkins Library
University of North Carolina at Charlotte

Owen King
American Archive of Public Broadcasting
GBH Archives

Jamie Patrick-Burns
Highlights for Children

Best Practices Exchange (Un)Conference 2023
Athens, Georgia
June 12, 2023

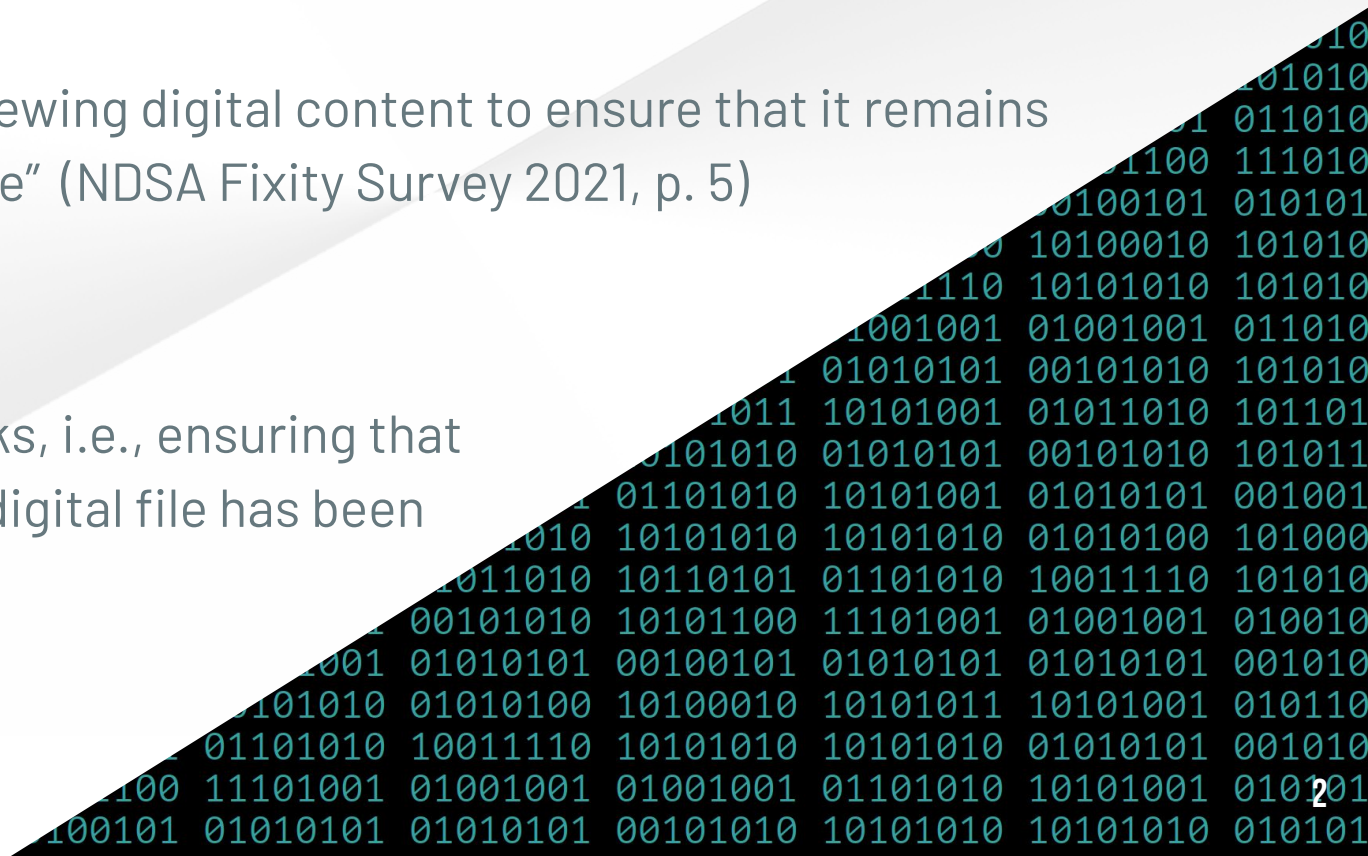
BIT-LEVEL FIXITY CHECKING

Fixity checking:

"the practice of reviewing digital content to ensure that it remains unchanged over time" (NDSA Fixity Survey 2021, p. 5)

Our focus:

Bit-level fixity checks, i.e., ensuring that not a single bit in a digital file has been altered.



FIXITY CHECKS AT THE STATE ARCHIVES OF NORTH CAROLINA

Archives Servers

- SHA256 checksums generated and checked with Bagger or bagit-python
- Run on new ingests, after transfer, and annually

LibSafe

- Checksums validated on ingest with Bagger integration
- Rolling verification

NC Dept of Natural & Cultural Resources Archives & History Building. Taken August 19, 2009. Public domain. [https://commons.wikimedia.org/wiki/File:D_2009_9_255_Archives_and_History-State_Library_building_8-19-09.jpg]

FIXITY CHECKS AT J. MURREY ATKINS LIBRARY UNC CHARLOTTE

Forensic Disk Images

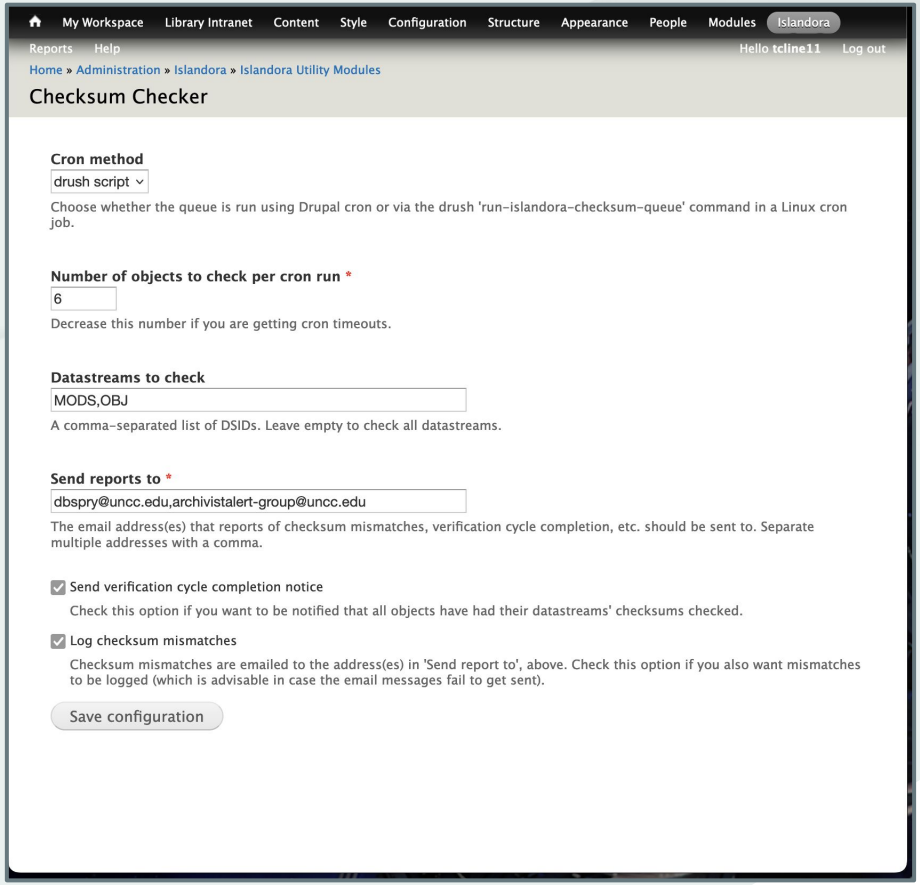
Checksums generated with Guymager

- MD5, SHA-1, and SHA256
- Fixity checked with Windows 10 command line or MD5Summer

Islandora Digital Objects

Checksums generated by Islandora

- BagIt generates bags from objects on ingest
- SHA-1 (options for MD5, SHA256)
- Checked with scripts or Bagger
- Processor-intensive on EC2 Server



THE WIDESPREAD PRACTICE OF FIXITY CHECKING

Data from the 2021 NDSA Fixity Survey
(116 respondents from across the global digital preservation community)

"Do your institutional practices include utilizing fixity information at any point in time?"

→ **97% answered YES.**

"What types of fixity information do you employ on files you are managing for the long-term?"

→ **98% included "Checksums and/or hash values"**

"For data at rest (i.e. in storage) do you check fixity information at regular time-based intervals?"

→ **71% gave a positive answer.**

BUT WHAT DO WE KNOW ABOUT THE VALUE OF FIXITY CHECKING ???

NOT NEARLY ENOUGH.

Results from the NDSA Survey

26% of survey respondents employed different fixity practices on different types of storage media.

73.2% of respondents experienced fixity failures at some point in time.

Other sources of info on disk failure

Differences between archival and IT/computing perspectives

- Disk block level vs. file level integrity
- Errors in transfer vs. bit rot
- Short-term failures of online storage vs. long-term fixity failures.

SO, OUR QUESTION: HOW CLOSELY TO WATCH THE BITS?

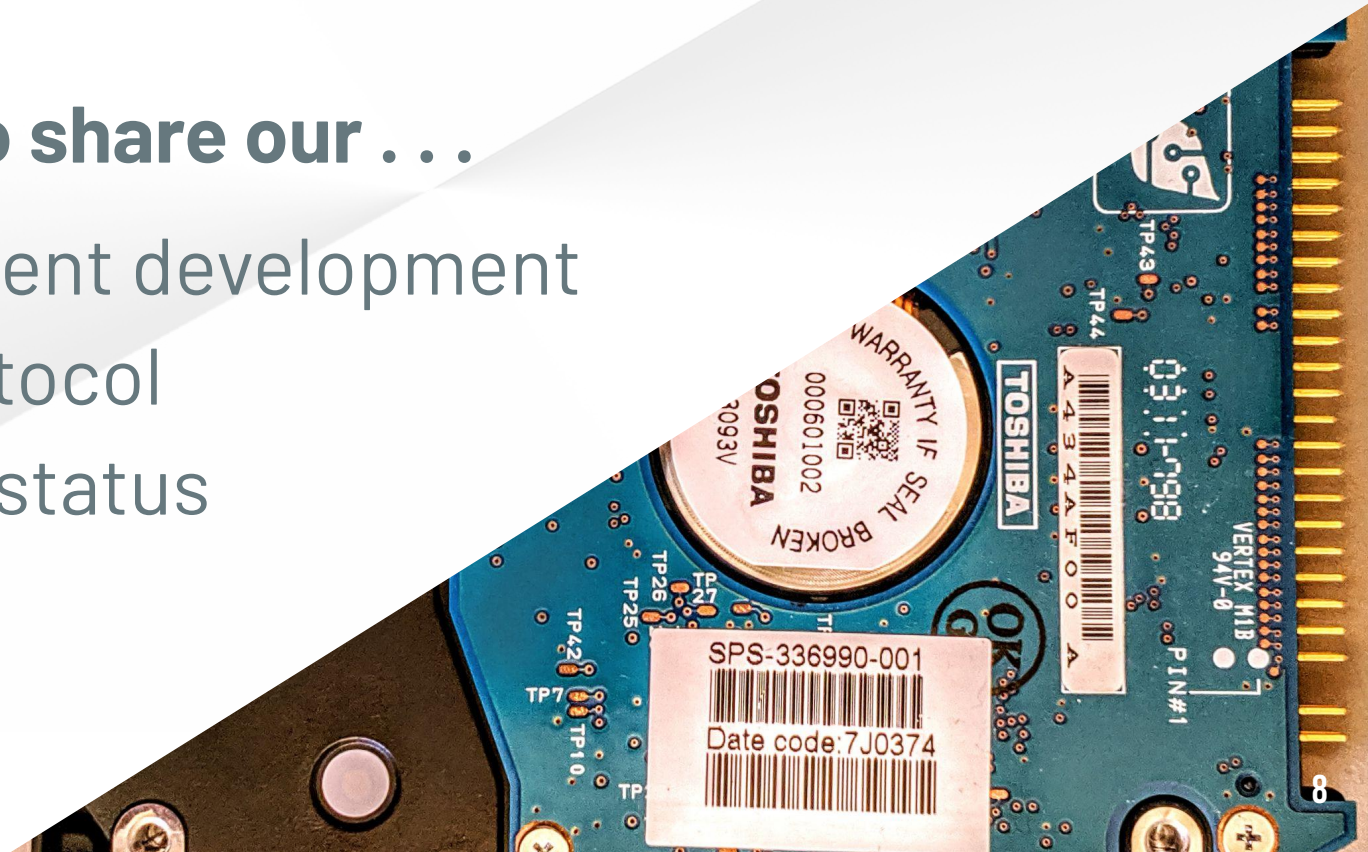
"we have allowed techniques appropriate to a different age to survive unchallenged in an era dominated by collection materials that are profoundly different in both volume and character." (MPLP)

"Because there is no point in preserving digital content if there will be no future generation of users, responsible digital preservation programs will seek to reduce carbon outputs. . . While cultural heritage organizations rarely seek to make a profit, economic sustainability is vital to organizational health and costs for digital preservation must be controlled." (Tallman, 2021)

THE EXPERIMENT

We'd like to share our . . .

- experiment development
- test protocol
- current status



DEVELOPING THE EXPERIMENT

Defining the Scope

What kind of storage media do we want to test? What kind of data do we want to store on it? Over what period of time?

Large files, small files, differing file types, random data?

Determination

The test is to look for failures down to the bit level, so we're only testing the bitstreams, not any particular file types

Payload Content

A set of files around 45 MB each, with random data, stored up to the capacity of the storage medium

CHOICE OF MEDIA TO FOCUS ON

Brainstorming Session for Media Types
Magnetic, optical, solid-state, cloud?

So Many Subtypes
SMR/CMR HDD, CD/DVD/BD,
SLC/TLC/QLC NAND, etc.

Winnowing the List
Not testing cloud (just storage on someone else's computer)
Not testing most solid-state (expense per GB)
Focus on CMR hard drives and some CDs/DVDs (ubiquity and affordability)

OVERVIEW OF THE TEST PROTOCOL

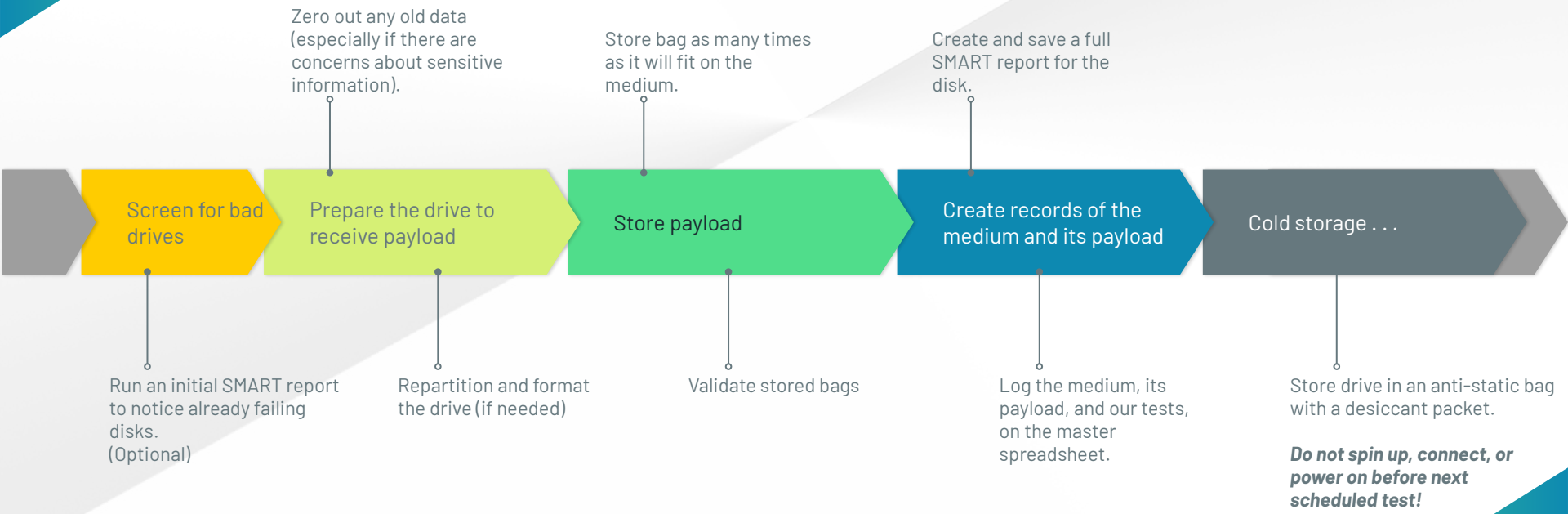
Basic idea
Store a payload on a disk, and check the fixity years later!

Tools
Use SMART data (via smartmontools) to record drive health.
BagIt with SHA-256 hashes for fixity checking.

Method
Use standardized payloads that can be compared to reference copies in case of fixity failures.

TEST PROTOCOL DETAILS

We keep a document that lists the steps (and relevant shell commands) to follow for each new drive we add to the experiment.

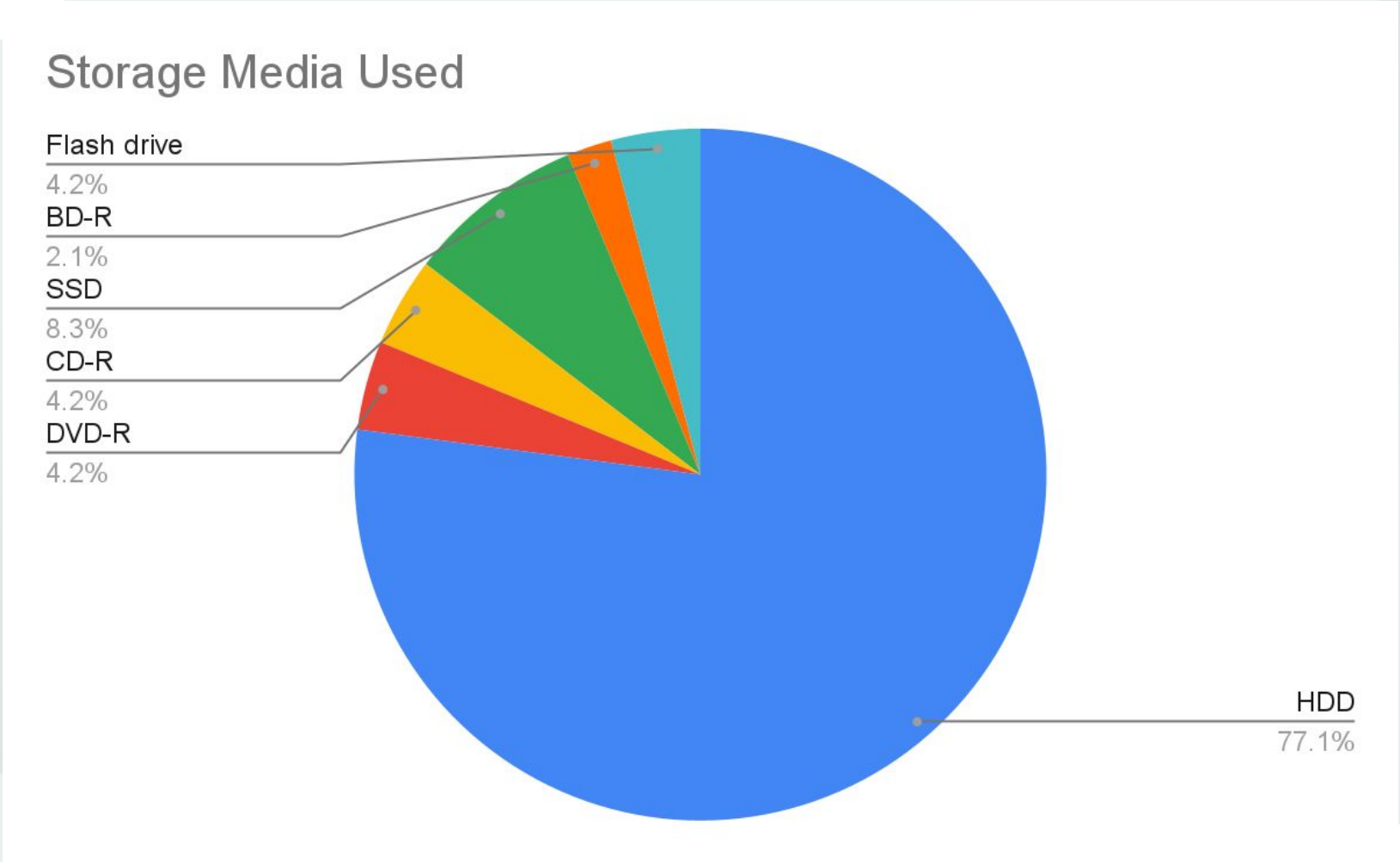


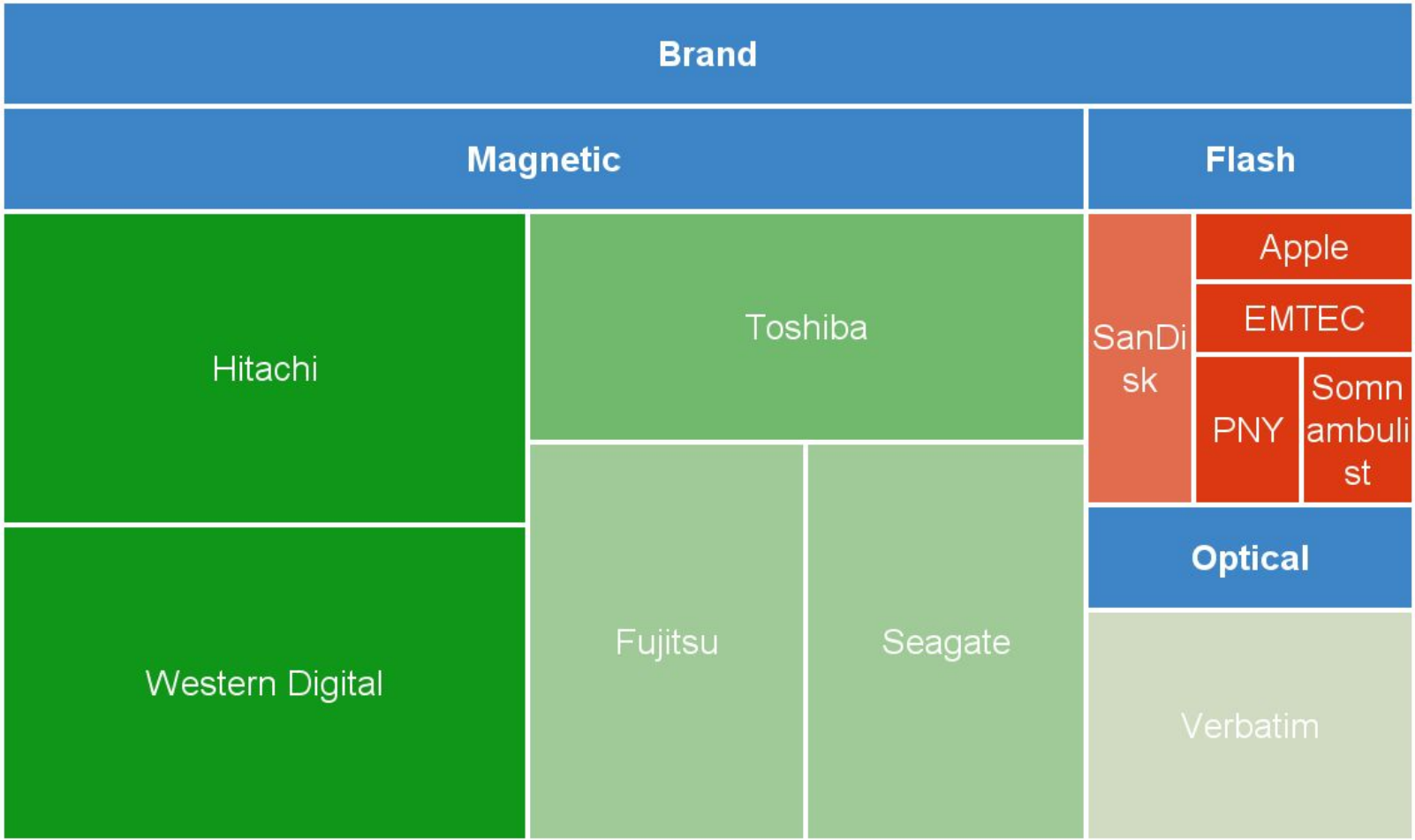
TEST STATUS

The Initial Test Set

- Tests initiated November 2022 - February 2023
- 48 media, 163 bags, 8.53 TB of data

Storage Type	Count
Magnetic	37
Flash	6
Optical	5
Total	48


Storage Media Used
Flash drive
4.2%
BD-R
2.1%
SSD
8.3%
CD-R
4.2%
DVD-R
4.2%
HDD
77.1%


Brand
Magnetic
Flash
Hitachi
Toshiba
SanDisk
Apple
EMTEC
PNY
Somnambulist
Western Digital
Fujitsu
Seagate
Optical
Verbatim

CAN WE HEAR FROM YOU?

The focus of our experiment is the fixity of data at rest.

Our questions for you:

(1) Do any of your procedures include fixity checks?

(2) Is it your policy to check the fixity of data at rest?

(3) Have you ever found a fixity failure in your data at rest?

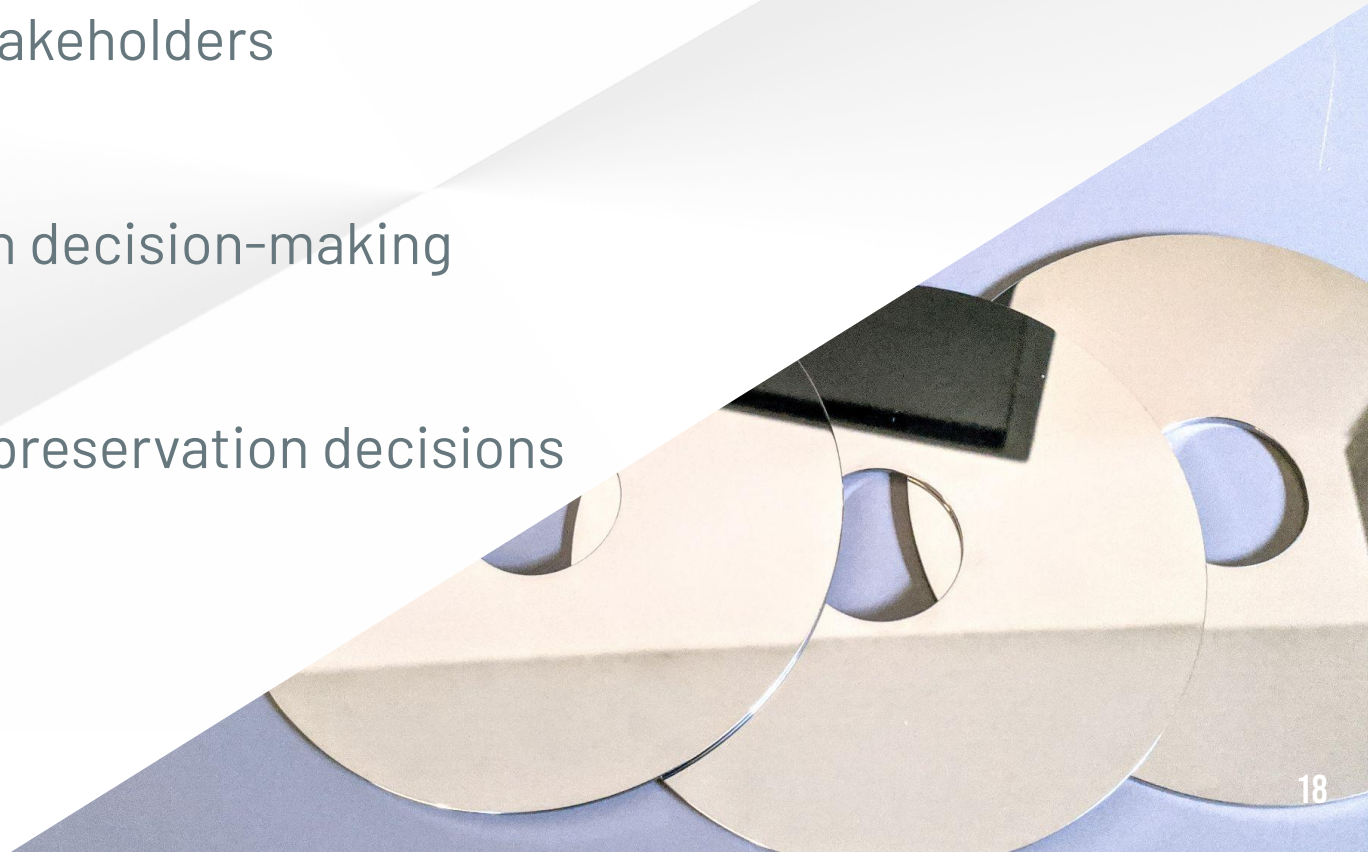
SO, MIGHT WE BE WASTING TIME ON FIXITY CHECKS?

Obviously, we do not know how this experiment will turn out.

If it turns out that the vast majority of magnetic media is relatively shelf stable for, say, 10 years, **would that change your preservation priorities**?

PAYOFFS OF EXPERIMENTATION

- Advising stakeholders
- Data-driven decision-making
- Justifying preservation decisions



REFERENCES

Greene, M. and Meissner, D., 2005. More product, less process: Revamping traditional archival processing. The American Archivist, 68(2), pp.208-263. https://doi.org/10.17723/aarc.68.2. c741823776k65863

NDSA Fixity Survey Working Group. 2021. "Results of the 2021 Fixity Survey." https://osf.io/2qkea/.

Tallman, Nathan. 2021. "A 21st Century Technical Infrastructure for Digital Preservation". *Information Technology and Libraries* 40 (4). https://doi.org/10.6017/ital.v40i4.13355.

Thank you!

We welcome your feedback and your participation!

If you have media to add to this experiment, please get in touch. We can arrange to take custody of the media, or share our documentation so you can collaborate with us.

Visit the **Bit Rot Fixity eXperiment** website: https://brfx.org
Get in touch with Tyler, Owen, and Jamie by emailing club@brfx.org.