TIPPING POINT

Have we gone past the point where we can handle the Digital Preservation Deluge?

Paul Stokes
Jisc
UK
paul.stokes@jisc.ac.uk
0000-0002-7333-4998

Karen Colbron
Jisc
UK
karen.colbron@jisc.ac.uk
0000-0002-2438-4008

Abstract – The world today is faced with an insurmountable problem. There is too much digital "stuff" in existence for us to even handle in any sort of meaningful way, let alone curate and preserve. We have reached (or perhaps even gone beyond) the data processing tipping point. There is an enormous amount of data already in existence and unimaginably more being generated every day. This panel proposal (and accompanying poster) is intended to explore this doomsday data scenario with a group of experts in the field of Digital Preservation and related disciplines with a view to deciding if it is true and what can be done about it.

Keywords – Data doomsday, Tipping point, Data deluge

Conference Topics – SUSTAINABILITY: REAL AND IMAGINED;. WE'RE ALL IN THIS TOGETHER; IMMERSIVE INFORMATION

I. INTRODUCTION

Some facts about the data-verse we currently inhabit.

There is an enormous amount of data existing in the world today. According to the International Data Corporation (IDC), the amount of digital data created, captured, and replicated in 2020 was approximately 64.2 zettabytes (1 zettabyte = 1 trillion gigabytes)[1]. This figure is expected to grow to 181 zettabytes by 2025, which represents a compound annual growth rate of 23%[1]. It is interesting to note that this figure is constantly being revised upwards. Publications from as recently as the late 2010's had this figure estimated as just over half that figure.

A report/infographic from the World Economic Forum based on data from by Seagate and IDC found that the amount of data generated each day is expected to reach 463 exabytes (1 exabyte = 1 billion gigabytes) by 2025, up from 23 exabytes in 2018[2]. This represents a CAGR of 29.4% over the seven-year period. The rate of world production of digital data is increasing at an extremely fast pace, with the amount of data generated each year growing by tens (possibly even hundreds in the future) of percent.

Why is this something to be worried about? Well generating the data is just beginning of the potential problem. The data needs to be transported, copied, and stored (and in some cases, curated and preserved), all of which require resources (including power) that are finite… and not keeping pace.

More facts.

As of 2022, it was estimated that the world had approximately 6.7 zettabytes (ZB) of data storage capacity and that this would rise to around 16 zettabytes by 2025 according to statistia[3]. This estimate includes all types of data storage, including hard disk drives, solid-state drives, optical storage, and tape storage. This number is constantly increasing, but not at the same pace as data production. Yes, there are newer and denser storage media on the horizon (or even in production) such as

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121078

DNA and storage on atoms, but even that is finite. An estimate published in the Straits Times gave a figure of about 180 years before all the atoms on earth were used for storage at the current rate of data production[4]. This is quite clearly a nonsense scenario, but it gives a flavour of the problem.

And what about manipulating the data. Ingesting it into a digital preservation system for instance. Doing it manually at high speed and large volumes is out of the question[*]. Semi-automated ingest (possibly enhanced by AI) should be faster. However, even this style of ingest is clearly several orders of magnitude slower than the rate at which data is being produced. Anecdotally, a growing number of data stewards are reporting that they are receiving data deposits faster than they can process them. As a result, the unprocessed data is being put into (at best) bit preserved long-term storage for processing "at a later date…" a later date that keeps moving further away. Automated ingest helps, but is still not going to keep pace.

So have we reached the tipping point? Are we past the point at which we can meaningfully process the deluge of data being generated? Is there anything that can be done to mitigate against data doomsday? That's what we'd like to explore in this panel.

II. The panel

We see this as an interactive panel session utilizing the expertise of up to 7 authorities from the field of Digital Preservation and related disciplines. Each will be asked to briefly put forward a point of view / opinion relating to the veracity (or otherwise) of the data doomsday scenario. The statements will be followed by a series of questions designed to explore how the situation could be either avoided or mitigated.

The audience will be invited to participate through a series of interactive polls and questions/observations from the floor (both those in attendance and those attending virtually). In particular, the audience will be polled at the beginning and the end of the session to see if they have been persuaded to shift their pre-extant opinions regarding the data doomsday scenario by the arguments presented in the session.

The following panelists have expressed a willingness to take part:

- William Kilbride—Executive Director of the DPC
- Matthew Addis—Chief Technology Officer at Arkivum
- Stephen Abrams—Head of Digital Preservation at Harvard University
- Kate Murray—Digital Projects Coordinator at Library of Congress
- Nancy McGovern—Director of Digital Preservation at MIT Libraries
- Tim Gollins—Director of Vanderbilt University Special Collections and University Archives
- Helen Hockx-Yu—Enterprise Architect at University of Notre Dame

1. REFERENCES

[1] International Data Corporation (IDC). (2020). The digitization of the world from edge to core. https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf

[2] https://www.weforum.org/agenda/2019/04/how-much-data-is-generated-each-day-cf4bddf29f/!

[3] Total installed based of data storage capacity in the global datasphere from 2020 to 2025, https://www.statista.com/statistics/1185900/worldwide-datasphere-storage-capacity-installed-base/

[4] https://www.straitstimes.com/singapore/world-faces-data-storage-crunch-ahead

[*] A moments consideration of the wide number and variety of processes and systems involved in the ingest process (multiple carrier types, multiple file types, multiple processes on multiple different infrastructures) leads us to the conclusion that a definitive, quantitative measure of manual ingest rate is impractical.