DNA Data storage for long term digital preservation

| **Euan Cochrane** | **Daniel Chadash** |
| --- | --- |
| *Yale University Library* *United States**euan.cochrane@yale.edu**0000-0001-9772-9743* | *Twist Bioscience**United States**dchadash@twistbioscience.com**0000-0002-9712-6034* |
|  |  |

**Abstract –**

**Long term digital preservation can benefit from long term storage solutions. DNA data storage is a new technology that offers unique benefits to the digital preservation community. With the cost of DNA data storage rapidly decreasing Yale University Library has partnered with Twist Bioscience to investigate the benefits and feasibility of deploying DNA data storage for long term preservation. In this paper we discuss DNA data storage, outline the pilot project we have undertaken and discuss the technology and its potential future applications.**

**Keywords – DNA, Storage**

**Conference Topics – Innovation, environment.**

# Introduction

In our increasingly volatile geopolitical world those of us working to preserve the cultural, scientific, and spiritual knowledge and records of our global societies in digital form have increasing justification to look for methods of storing these digital artifacts in ways that would outlast long periods of environmental and geopolitical volatility. DeoxyriboNucleic Acid (DNA) data storage technology provides an increasingly practical, sustainable, durable, and cost-effective method for mitigating many of the potential risks to digital storage infrastructure that seem increasingly likely to become realized as destructive issues in upcoming years.

In addition to DNA data storage being a reliable, durable and long lasting medium, media diversity is very important in any long-storage digital preservation strategy that can significantly de-risk the chances of data loss.

DNA data storage in its current form is a type of ‘Write Once, Read Forever’ digital storage media (WORF) [1]. WORF media is so-called because it can only be written once (cannot be edited) and once written will survive and be accessible ‘forever’ or at least for a hugely greater time than standard digital storage media. Other digital storage media described in the same way include:

1. M-DISC optical media - with a claimed lifespan of ‘1000 years’ [2]
2. PIQL Film technology - Film-based storage using binary data stored as Q R codes on the film, with an expected lifetime of over 500 years [3].
3. Digital Optical Technology System (DOTS) - “DOTS physically encodes data on an archival tape coated in a phase-change alloy” and has a claimed lifetime of over 100 years [4].

Using DNA data storage for long term bit-preservation is particularly attractive in the context of these other available options for a number of reasons. Specifically DNA data storage has unique qualities that make it attractive compared to these alternatives, including:

1. While future post-disaster humans may never re-develop (for example) a blu-ray reader, no matter how significant the disaster that could befall humanity, surviving humans ought to eventually want and need to read their DNA once more. In doing so this will provide the ability to unlock a potential deluge of valuable knowledge from the past.
2. DNA data storage is extremely information dense. This makes it very suitable for use in time capsules and data caches that could be deliberately hidden to prevent their destruction in times of volatility. Furthermore the relatively tiny [Figure 1] physical size of current DNA data storage containers make them less likely to be affected by any sort of physical event simply due to their minimal attack surface.



Figure 1: a DNA data storage capsule

Image courtesy of Imagene SA

The small physical size also means that the cost of dispersing copies of data stored as DNA is relatively low. This may potentially enable more varied and effective storage-risk mitigation strategies to be implemented at the same cost as it would be to implement a higher risk strategy using larger, more expensive media.

1. DNA data storage is increasingly cost-effective. Twist predicts that the cost will reduce to less than $1000 per terabyte within the next few years and the cost could be further reduced in the future. Equally relevant, once a set of data has been synthesized, the cost of additional copies of the dataset is very insignificant and mostly dependent mostly upon the cost of the (inexpensive - under tens of dollars each) physical container used to encapsulate the DNA material and the low cost of reagents to copy the DNA. This has wide implications for storage-risk mitigation strategies as it allows for significant physical redundancy to be implemented in storage strategies that use DNA data storage. For example, many additional copies of the same data may be able to be placed in locations with wide variations in their risk profiles at minimal additional cost.
2. DNA data storage has significant redundancy built into it which provides the opportunity for very effective error correction functionality to be included in the implementation of the data-to-DNA synthesis and recovery/reading processes. This means that it is far less likely that a single ‘copy’ of data stored as DNA will not be able to be read at any point in the future[[1]](#footnote-1).
3. Unlike some alternative options[[2]](#footnote-2) DNA has already been proven to last for a very long time in the real world [5]. In addition, existing examples of readable DNA samples have been recovered from physical contexts that are extremely far from the ideals that can be achieved using the encapsulation technologies we have available today. Such technologies coupled with a deliberate plan for locating the physical media in many low-risk locations make data stored as DNA extremely likely to be accessible in the future.
4. Environmental Benefits
	1. DNA data storage only uses energy when being read or written rest. At rest it requires no energy consumption It can be stored at room temperature and has no no active cooling requirements.
	2. Physically little material is necessary to store large volumes of data. A room of DNA would likely easily store all the data that exists in the world in 2022.
	3. The technology consists of very easily recyclable components. Just metal and organic material are required and the capsules can also be washed and reused.
	4. The sourcing of materials is easier than alternatives. For example there is no need for rare metals that might necessarily be sourced from conflict zones.
5. Offline storage benefits - DNA data storage can be kept offline (disconnected from the public internet or local intranet). This enables a higher level of security as for most risk profiles it prevents various threats such as ransomware attacks.

# Value in Long Term Digital Preservation

WORF media has a long history in digital preservation[[3]](#footnote-3). The community has never fully embraced it for a number of previously justifiable reasons. Amongst those reasons are:

1. The need to be able to make changes to preserved data or its metadata over time.
2. The need to check datasets for integrity over time, and the perceived need to continually change the digital objects over time so they work in current technologies (i.e. to undertake migration of content between files of different formats).
3. Digital preservation also continues to be a relatively nascent field that has not been around long enough to verify vendor claims of longevity. This has meant a natural lack of trust exists in unproven solutions
4. From an historic perspective the nascent digital preservation community has been operating in a period of relative global stability, especially in the western world in which most of the practitioners operate. WORF media is most easily justifiable as a hedge against loss due to instability. Without the risk of such threats it has been harder to justify the cost of mitigating against them by using WORF solutions.

For these reasons, to date, the digital preservation community has not had any compelling justification for implementing WORF media. However this may be changing.

Emulation solutions have become more widely used [7] and provide a novel solution for ensuring future generations can decode digital files into meaningful information. As opposed to migration, as a solution for ensuring long term access to content in preserved digital files, emulation can be implemented in a way that requires minimal or no change to stored digital files over time. Emulators, and the full computing environments needed to access stored digital files, can be stored using the same mechanisms as the digital files themselves (i.e. in this case using DNA data storage). Over time if an emulator becomes incompatible with current technology a new one can be written to replace the old emulator, and stored alongside the old ones, or a new emulator can be created into which the old one can be nested. In either case, the primary digital objects being preserved do not need to be altered at all, making write-once media feasible.

## Long Term Preservation Challenges of DNA Data Storage.

As discussed, emulation offers a practical option for ensuring future users can access digital objects stored using DNA data storage without requiring the digital objects be regularly replaced. However neither emulation nor migration can mitigate against large gaps in time between storing objects and future users trying to access them, during which time the computing technology (and related knowledge) required may have been lost. For such a scenario we will need additional infrastructure in place to boot-strap future generations to the point where they can make sense of binary data (for example there is an attempt to create a “Manual for [rebuilding] civilization”[8]).

# Turning DNA Data Storage Into a Write-Many Service

While this article has focussed on the use of DNA data storage as a WORF technology Twist has plans in place to offer an implementation of the technology that would make it functionally similar to re-writable media. In the planned configuration DNA data storage will serve a similar role to archivally-configured cloud storage like AWS Deep Glacier [9] and can offer a new cold layer in the archival tiers.

Current large scale tape based digital storage solutions generally involve the use of a central management server with local higher speed cache storage, one or more tape drives, a large storage rack to store tape cartridges, and a ‘tape robot’ that can fetch cartridges and insert them in the drive to be read and their data cached to the local server for end-user access. A similar approach is used for large scale blu-ray storage solutions [10]. The difference between those two however is that blu-ray disks are generally implemented as write-once media meaning edits are not possible and deletions require destroying entire discs. The blu-ray based approach is currently the most similar to that which is planned for the future utilizing DNA data storage. Using the new approach it will be possible to implement automated machines to synthesize data as DNA, and to read the data back from the synthesized DNA before re-synthesising the data with changes and storing the DNA back in it’s storage containers. Implementing a process like this will effectively turn DNA data storage into a practical automated medium/long term storage mechanism.

Unfortunately the current cost of synthesizing and reading DNA data are both too high for practical use in this way [11]. In addition the time taken for both activities is prohibitively high such that it would not be practical for usage in active-archive scenarios where access time and frequency are important. However it might be within the acceptable range for becoming a competitor for offline tape storage or whatever technology is backing the deep-archive solutions offered by cloud vendors such as Amazon’s ‘Glacier Deep Archive’ , which currently only guarantees retrieval within 12 hours.

An even simpler potential use for DNA data storage for shorter term (less than ‘forever’) storage requirements will arise if the cost for synthesis/storage reduces even more dramatically. If the cost of DNA data storage reaches a level significantly less than that of other options it will become viable to use it as an additional risk-mitigating back-up copy. Used in this way, edits to stored data would simply involve deletions and inexpensive re-encodings of changed data, and partial-deletions would be similar.

# The DNA data storage Pilot Project at Yale University Library

The Yale University Digital Preservation unit provides digital preservation services across the libraries, archives and museums on campus. The services are used to preserve a huge variety of collections ranging from the Forunoff Holocaust Testimonies, to architecture records, to video games, and more [cite]. Most of Yale’s preserved collections are unique and irreplaceable and as such are important to the global historic record for informing future generations. The Digital Preservation Services (DPS) team is organized as an internal service provider and works with collection owners to provide services so that they can preserve their content. As part of these services DPS provides multiple options for storing digital content in ways that have different risk profiles associated with them (different sets of specific risks that may lead to data being lost, with different likelihoods associated with them). DPS does not decide the storage options that are used as part of the bit preservation strategies employed to preserve Yale’s collections. Instead DPS discusses the options with collection owners and they decide what option best fits their risk appetites and budgets.

As part of providing these services the DPS team has recently been investigating options for offering WORF media offerings to its users. The DPS team expects that at least initially only the most high-value collections will find these options attractive due to the cost, however this may be naive, especially if costs continue to decrease as predicted by some service providers.

In 2021 the DPS team began a pilot project to test the DNA data storage technology offered by Twist Bioscience. The team’s goals in the pilot were:

1. To learn about the technology, e.g. how it works, how feasible it is to implement.
2. To test the technology over time
3. To socialize the idea of DNA data storage and seek feedback from our stakeholders about their interest in implementing it
4. To use this interesting new technology to raise awareness about digital preservation in general.

The pilot involved storing approximately 15 megabytes of data at a cost of nearly $1000 per Megabyte. The DPS team worked with librarians at the Harvey Cushing/John Hay Whitney Medical Library to select sample data that was open, relevant, and interesting. The hope was to include data that could be publicly shared in publications used for awareness raising about the project and which would attract interest.

The data was packaged into a single .zip file as the current implementation of the DNA data storage technology does not include a filesystem, preventing multiple files being stored in the same DNA sequence. Following this the data was passed on to Twist Bioscience (Twist) to be converted from binary to a DNA sequence. After this the DNA sequence was synthesized, replicated, and stored in 40 capsules.

The DPS team requested 40 capsules for two purposes. Firstly the team intends on undertaking test-reads of the data in the capsules at regular intervals, and so needed multiple copies to support this. Secondly some of the capsules will be given away as keepsakes as part of awareness raising activities.

The DNA synthesis process was relatively short and the DPS team received the capsules with the data within a few weeks of sending off the file for synthesis. Following this the team reached out to researchers in the Yale Center for Genome Analysis (YCGA) who offered to conduct an initial reading of the DNA in one of the capsules. This process was undertaken using standard DNA sequencing technology that the YCGA researchers use on a regular basis. After sequencing the DNA the resulting DNA sequence had to be converted back to binary data. This process was undertaken using a virtual machine provided by Twist, containing the decoding algorithm and software implementation. Converting the sequence to binary resulted in the original file that had the original checksum associated with it. The DNA Data Storage Alliance is planning to develop an industry standard encoding and decoding algorithm and implementation. These are intended to be open source and freely available.

The DPS team intends to replicate this reading process in (at a minimum) 2, 5, and 10 years time and report the results publicly in order to provide a set of benchmark data for others in the digital preservation community to learn from. We (Twist and the DPS team) are also planning to undertake a second pilot in 2022 to take advantage of the new iteration of the storage technology that has an estimated cost of US$1000/GB, which is an improvement of three orders of magnitude of the cost.

# The DNA Data Storage Technology

Storing data on DNA is not a new concept. It was first demonstrated a few decades ago.

The dominant workflow today for storing and reading data on DNA is built from 6 steps:



Figure 2: The DNA Data Storage Workflow

1. **Coding** - Any digital file at its base is built from 1s and 0s, using an encoder we translate the 1s and 0s into AGTC (the building block of DNA), and divide the long string of letters into short sequences so it will fit current synthesis technologies. In addition we add metadata and error correction codes to help us with reading the data back in case of sequencing errors.
2. **Synthesis** - The short sequences are sent to synthesis, where we translate the text strings of ATGC into real physical DNA using a DNA “printer”.
3. **Storage** - The DNA comes out of synthesis in liquid form, it is dehydrated for long-term storage purposes and is encapsulated and sealed in capsules for storage.
4. **Retrieval** - Once access to the data is needed, the capsule is opened and the DNA is rehydrated and prepared for sequencing/reading.
5. **Sequencing** - The liquid DNA is read using a device called sequencer. The sequencer transforms the physical sample into a digital form of the short strings of DNA letters.
6. **Decoding** - Using the codec that was used in step 1, the decoder is able to reconstruct the digital binary file from the short DNA sequences using the metadata in each short string and is able to correct errors that originated from the sequencing process.

This workflow is mainly deployed in demonstrations and proof of concepts of the technology through the past decade mainly because of the cost of DNA synthesis, as there was no new technology to enable a drastic cost reduction.

Twist Bioscience was founded in 2012 with a breakthrough technology for synthesizing DNA. Using the advancements in chip and silicone manufacturing Twist was able to synthesize DNA on a silicone chip that allowed miniaturization of the components that enabled the lowest price point to date for custom DNA.
A new generation of that original chip was used for the first part of the pilot at Yale University Library which enabled the first commercial offering of DNA Data Storage at $1000/MB.

In order to reach commercial viability, there still needs to be an improvement of a few orders of magnitude of the cost.

## Scaling the Technology

The technology that was used in the first part of the pilot is the same one that is being used to synthesize DNA for a broad range of applications in the healthcare and biotechnology industries. Very high quality DNA, without many errors that is suited for applications that can’t tolerate a moderate error rate.

The advantage of using DNA for data storage is that we can incorporate many well-proven algorithms from information/communication theory that allows us to recover the data even if there are errors in the synthesis/sequencing processes, in the same way that our cellphones are able to compensate on noise/error in the cell signal.

This property allows us to develop new technologies that are specific for DNA data storage and will allow a massive cost reduction. Most of the cost reduction will come from miniaturizing the features and building more reagent-efficient chips and technologies.

The second chip (POC chip) that will be used for the second part of the pilot is a DNA data storage dedicated chip that was developed by Twist specifically for that purpose and shows a staggering improvement of three orders of magnitude in the cost (from $1000/MB to ~$1000/GB).

As described in the figure below, Twist is not stopping at that price point and plans to continue with the cost reduction until it will be able to compete with prices of current archival storage mediums (tape and HDD).



Figure 3: The DNA data storage synthesis scaling innovation roadmap.

Without synthesis nothing further would be possible with DNA as a data storage technology. However it’s worth noting that there are more challenges down the road, and we focused here mainly on the synthesis as it’s the main enabler for the successful large scale use of DNA as a long term storage technology.

In addition to lowering the cost of synthesis there needs to be a drastic reduction in the cost of sequencing (few orders of magnitude) as without it customers won’t adopt the technology outside as a last-resort backup option.

Automation and scalability are also very important factors. Any technology that would be deployed in the data center needs to be highly automated and provide the same ease-of-use as current technologies and important features that other technologies provide today for archival applications.

#

# Conclusion

DNA data storage offers an increasing practical option for the long term storage of digital data. The successful pilot project trialing DNA data storage at Yale University offers a promising window into this new technology. With costs consistently decreasing DNA and new implementations in development that will turn DNA data storage technology into an effectively write-many media, we expect to see an increasing uptake in its use for long term preservation storage.

# REFERENCES

1. *Write Once, Read Forever (WORF)—Proof-Of-Concept Demonstrated For Archival Data Storage Using Interference Spectra*. Soloman. R. et al. Archiving 2015 Final Program and Proceedings. 2015. <https://www.ingentaconnect.com/contentone/ist/ac/2015/00002015/00000001/art00023>
2. *Accelerated Life Cycle Comparison of Millenniata Archival DVD*. I. Svrcek. Naval Air Warfare Center Weapons Division. 2009. [https://web.archive.org/web/20210211221229/https://www.esystor.com/images/China\_Lake\_Full\_Report.pdf](https://web.archive.org/web/20210211221229/https%3A//www.esystor.com/images/China_Lake_Full_Report.pdf) l
3. *PIQL Technology website*. PIQL. Retrieved 6th March 2022. <https://www.piql.com/about/technology/>
4. *DOTS—Long-Term, Human-Readable Archival Data Storage*. Warner. A. The Long Now Foundation. 2015. [https://web.archive.org/web/20220123123847/https://longnow.org/ideas/02015/12/27/dots-long-term-human-readable-archival-data-storage/](https://web.archive.org/web/20220123123847/https%3A//longnow.org/ideas/02015/12/27/dots-long-term-human-readable-archival-data-storage/)
5. *Million-year-old DNA sheds light on the genomic history of mammoths*. van der Valk, T., Pečnerová, P., Díez-del-Molino, D. et al. Nature 591, 265–269 (2021). <https://doi.org/10.1038/s41586-021-03224-9>
6. *Digital-Imaging and Optical Digital Data Disk Storage Systems*, The Technology Research Staff. The National Archives at College Park. 1994. Retrieved 6th March 2022. <https://www.archives.gov/preservation/technical/imaging-storage-appendix.html>
7. *Emulation as a Service Infrastructure (EaaSI)*. EaaSI team. 2018. Retrieved 6th March 2022. <https://www.softwarepreservationnetwork.org/emulation-as-a-service-infrastructure/>
8. *How Can We Create a Manual For Civilization?*. A. Kabil. The Long Now Foundation. 2017 <https://longnow.org/ideas/02017/06/07/how-can-we-create-a-manual-for-civilization/>
9. *New Amazon S3 Storage Class – Glacier Deep Archive*. J. Barr. Amazon Web Services. 2019. Retrieved 6th March 2022 <https://aws.amazon.com/blogs/aws/new-amazon-s3-storage-class-glacier-deep-archive/>
10. *Inside Facebook’s Blu-Ray Cold Storage Data Center*. R. Miller. Data Center Frontier, 2015. Retrieved 6th March 2022. <https://datacenterfrontier.com/inside-facebooks-blu-ray-cold-storage-data-center/>
11. *Preserving our digital legacy: an introduction to DNA data storag*e. The DNA data storage alliance. 2021. Retrieved 6th March 2022. <https://dnastoragealliance.org/dev/wp-content/uploads/2021/06/DNA-Data-Storage-Alliance-An-Introduction-to-DNA-Data-Storage.pdf>
1. ‘Copy’ is not entirely accurate here as each DNA data storage capsule includes a huge number of copies of the data being stored. [↑](#footnote-ref-1)
2. Though not all - e.g. the PIQL technology uses time-proven film media for it’s physical storage mechanism. [↑](#footnote-ref-2)
3. See the summary section in “Site visit report #1” describing WORM use in 1988 in [6] [↑](#footnote-ref-3)