DNA digital data storage

**DNA digital data storage** is the process of encoding and decoding binary data to and from synthesized strands of [DNA](https://en.wikipedia.org/wiki/DNA).[[1]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Ceze_2018-1)[[2]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-pmid30073589-2)

While DNA as a storage medium has enormous potential because of its high storage density, its practical use is currently severely limited because of its high cost and very slow read and write times.[[3]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-pmid29744271-3)

In June 2019, scientists reported that all 16 GB of text from [Wikipedia's](https://en.wikipedia.org/wiki/Wikipedia) English-language version had been encoded into [synthetic DNA](https://en.wikipedia.org/wiki/Synthetic_DNA).[[4]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-CNET-20190629-4) In 2021, scientists reported that a custom DNA data writer had been developed that was capable of writing data into DNA at 18 Mbps.[[5]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-biorxiv.org-5)

Encoding methods[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=1)]

Countless methods for encoding data in DNA are possible. The optimal methods are those that make economical use of DNA and protect against errors.[[6]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Smith2003-6) If the message DNA is intended to be stored for a long period of time, for example, 1,000 years, it is also helpful if the sequence is obviously artificial and the reading frame is easy to identify.[[6]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Smith2003-6)

**Encoding text**[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=2)]

Several simple methods for encoding text have been proposed. Most of these involve translating each letter into a corresponding "codon", consisting of a unique small sequence of nucleotides in a [lookup table](https://en.wikipedia.org/wiki/Lookup_table). Some examples of these encoding schemes include [Huffman codes](https://en.wikipedia.org/wiki/Huffman_codes), comma codes, and alternating codes.[[6]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Smith2003-6)

**Encoding arbitrary data**[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=3)]

To encode arbitrary data in DNA, the data is typically first converted into [ternary](https://en.wikipedia.org/wiki/Ternary_numeral_system) (base 3) data rather than [binary](https://en.wikipedia.org/wiki/Binary_number) (base 2) data. Each digit (or "trit") is then converted to a nucleotide using a lookup table. To prevent homopolymers (repeating nucleotides), which can cause problems with accurate sequencing, the result of the lookup also depends on the preceding nucleotide. Using the example lookup table below, if the previous nucleotide in the sequence is T (thymine), and the trit is 2, the next nucleotide will be G (guanine).[[7]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Goldman2013-7)[[8]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Lee2019-8)

|  |  |  |  |
| --- | --- | --- | --- |
| **Trits to nucleotides (example)** | | | |
| **Previous** | **0** | **1** | **2** |
| **T** | A | C | G |
| **G** | T | A | C |
| **C** | G | T | A |
| **A** | C | G | T |

Various systems may be incorporated to partition and address the data, as well as to protect it from errors. One approach to error correction is to regularly intersperse synchronization nucleotides between the information-encoding nucleotides. These synchronization nucleotides can act as scaffolds when reconstructing the sequence from multiple overlapping strands.[[8]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Lee2019-8)

In vivo[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=4)]

The genetic code within living organisms can potentially be co-opted to store information. Furthermore [synthetic biology](https://en.wikipedia.org/wiki/Synthetic_biology) can be used to engineer cells with "molecular recorders" to allow the storage and retrieval of information stored in the cell's genetic material.[[1]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Ceze_2018-1) [CRISPR gene editing](https://en.wikipedia.org/wiki/CRISPR_gene_editing) can also be used to insert artificial DNA sequences into the genome of the cell.[[1]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Ceze_2018-1) For encoding developmental lineage data (molecular flight recorder), roughly 30 trillion cell nuclei per mouse \* 60 recording sites per nucleus \* 7-15 bits per site yields about 2 TeraBytes per mouse written (but only very selectively read).[[9]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-9)

History[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=5)]

|  |  |
| --- | --- |
|  | This section **relies excessively on**[**references**](https://en.wikipedia.org/wiki/Wikipedia:Verifiability)**to**[**primary sources**](https://en.wikipedia.org/wiki/Wikipedia:No_original_research#Primary,_secondary_and_tertiary_sources). Please improve this section by adding [secondary or tertiary sources](https://en.wikipedia.org/wiki/Wikipedia:No_original_research#Primary,_secondary_and_tertiary_sources). *Find sources:* ["DNA digital data storage"](https://www.google.com/search?as_eq=wikipedia&q=%22DNA+digital+data+storage%22) – [news](https://www.google.com/search?tbm=nws&q=%22DNA+digital+data+storage%22+-wikipedia&tbs=ar:1) **·** [newspapers](https://www.google.com/search?&q=%22DNA+digital+data+storage%22&tbs=bkt:s&tbm=bks) **·** [books](https://www.google.com/search?tbs=bks:1&q=%22DNA+digital+data+storage%22+-wikipedia) **·** [scholar](https://scholar.google.com/scholar?q=%22DNA+digital+data+storage%22) **·** [JSTOR](https://www.jstor.org/action/doBasicSearch?Query=%22DNA+digital+data+storage%22&acc=on&wc=on) *(April 2018) (*[*Learn how and when to remove this template message*](https://en.wikipedia.org/wiki/Help:Maintenance_template_removal)*)* |

The idea of DNA digital data storage dates back to 1959, when the physicist [Richard P. Feynman](https://en.wikipedia.org/wiki/Richard_Feynman), in "There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics" outlined the general prospects for the creation of artificial objects similar to objects of the microcosm (including biological) and having similar or even more extensive capabilities.[[10]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-10) In 1964–65, [Mikhail Samoilovich Neiman](https://en.wikipedia.org/wiki/Mikhail_Samoilovich_Neiman), the Soviet physicist, published 3 articles about microminiaturization in electronics at the molecular-atomic level, which independently presented general considerations and some calculations regarding the possibility of recording, storage, and retrieval of information on synthesized DNA and RNA molecules.[[11]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-11)[[12]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-12)[[13]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-13) After the publication of the first M.S. Neiman's paper and after receiving by Editor the manuscript of his second paper (January, the 8th, 1964, as indicated in that paper) the interview with cybernetician [Norbert Wiener](https://en.wikipedia.org/wiki/Norbert_Wiener) as published.[[14]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-14) N. Wiener expressed ideas about miniaturization of computer memory, close to the ideas, proposed by M. S. Neiman independently. These Wiener's ideas M. S. Neiman mentioned in the third of his papers. This story is described in details.[[15]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-15)

One of the earliest uses of DNA storage occurred in a 1988 collaboration between artist [Joe Davis](https://en.wikipedia.org/wiki/Joe_Davis_(artist)) and researchers from Harvard. The image, stored in a DNA sequence in *E.coli*, was organized in a 5 x 7 matrix that, once decoded, formed a picture of an ancient Germanic rune representing life and the female Earth. In the matrix, ones corresponded to dark pixels while zeros corresponded to light pixels.[[16]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-:1-16)

In 2007 a device was created at the University of Arizona using addressing molecules to encode mismatch sites within a DNA strand. These mismatches were then able to be read out by performing a restriction digest, thereby recovering the data.[[17]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-17)

In 2011, George Church, Sri Kosuri, and Yuan Gao carried out an experiment that would encode a 659-kb book that was co-authored by Church. To do this, the research team did a two-to-one correspondence where a binary zero was represented by either an adenine or cytosine and a binary one was represented by a guanine or thymine. After examination, 22 errors were found in the DNA.[[16]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-:1-16)

In 2012, [George Church](https://en.wikipedia.org/wiki/George_M._Church) and colleagues at [Harvard University](https://en.wikipedia.org/wiki/Harvard_University) published an article in which DNA was encoded with digital information that included an HTML draft of a 53,400 word book written by the lead researcher, eleven JPG images and one JavaScript program. Multiple copies for redundancy were added and 5.5 [petabits](https://en.wikipedia.org/wiki/Petabit) can be stored in each cubic millimeter of DNA.[[18]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-church-18) The researchers used a simple code where bits were mapped one-to-one with bases,[[*clarification needed*](https://en.wikipedia.org/wiki/Wikipedia:Please_clarify)] which had the shortcoming that it led to long runs of the same base, the sequencing of which is error-prone. This result showed that besides its other functions, DNA can also be another type of storage medium such as hard drives and magnetic tapes.[[19]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-nanews-19)

In 2013, an article led by researchers from the [European Bioinformatics Institute](https://en.wikipedia.org/wiki/European_Bioinformatics_Institute) (EBI) and submitted at around the same time as the paper of [Church](https://en.wikipedia.org/wiki/George_M._Church) and colleagues detailed the storage, retrieval, and reproduction of over five million bits of data. All the DNA files reproduced the information between 99.99% and 100% accuracy.[[20]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Goldman_2013-20) The main innovations in this research were the use of an error-correcting encoding scheme to ensure the extremely low data-loss rate, as well as the idea of encoding the data in a series of overlapping short [oligonucleotides](https://en.wikipedia.org/wiki/Oligonucleotide) identifiable through a sequence-based indexing scheme.[[19]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-nanews-19) Also, the sequences of the individual strands of DNA overlapped in such a way that each region of data was repeated four times to avoid errors. Two of these four strands were constructed backwards, also with the goal of eliminating errors.[[20]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Goldman_2013-20) The costs per megabyte were estimated at $12,400 to encode data and $220 for retrieval. However, it was noted that the exponential decrease in DNA synthesis and sequencing costs, if it continues into the future, should make the technology cost-effective for long-term data storage by 2023.[[19]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-nanews-19)

In 2013, a software called DNACloud was developed by Manish K. Gupta and co-workers to encode computer files to their DNA representation. It implements a memory efficiency version of the algorithm proposed by Goldman et al. to encode (and decode) data to DNA (.dnac files).[[21]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-21)[[22]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-22)

The long-term stability of data encoded in DNA was reported in February 2015, in an article by researchers from [ETH Zurich](https://en.wikipedia.org/wiki/ETH_Zurich). The team added redundancy via [Reed–Solomon error correction](https://en.wikipedia.org/wiki/Reed%E2%80%93Solomon_error_correction) coding and by encapsulating the DNA within silica glass spheres via [Sol-gel](https://en.wikipedia.org/wiki/Sol-gel) chemistry.[[23]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-grass-23)

In 2016 research by Church and [Technicolor Research and Innovation](https://en.wikipedia.org/wiki/Technicolor_SA) was published in which, 22 MB of a MPEG compressed movie sequence were stored and recovered from DNA. The recovery of the sequence was found to have zero errors.[[24]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Blawat-24)

In March 2017, [Yaniv Erlich](https://en.wikipedia.org/wiki/Yaniv_Erlich) and [Dina Zielinski](https://en.wikipedia.org/w/index.php?title=Dina_Zielinski&action=edit&redlink=1) of [Columbia University](https://en.wikipedia.org/wiki/Columbia_University) and the [New York Genome Center](https://en.wikipedia.org/wiki/New_York_Genome_Center) published a method known as DNA Fountain that stored data at a density of 215 petabytes per gram of DNA. The technique approaches the [Shannon capacity](https://en.wikipedia.org/wiki/Shannon_capacity) of DNA storage, achieving 85% of the theoretical limit. The method was not ready for large-scale use, as it costs $7000 to synthesize 2 megabytes of data and another $2000 to read it.[[25]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-25)[[26]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-ScienceNews2017-26)[[27]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Science2017-27)

In March 2018, [University of Washington](https://en.wikipedia.org/wiki/University_of_Washington) and [Microsoft](https://en.wikipedia.org/wiki/Microsoft) published results demonstrating storage and retrieval of approximately 200MB of data. The research also proposed and evaluated a method for [random access](https://en.wikipedia.org/wiki/Random_access) of data items stored in DNA.[[28]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Organick_Ang_Chen_Lopez_2018_p.-28)[[29]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Patel_2018-29) In March 2019, the same team announced they have demonstrated a fully automated system to encode and decode data in DNA.[[30]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-30)

Research published by [Eurecom](https://en.wikipedia.org/wiki/Eurecom" \o "Eurecom) and [Imperial College](https://en.wikipedia.org/wiki/Imperial_College_London) in January 2019, demonstrated the ability to store structured data in synthetic DNA. The research showed how to encode structured or, more specifically, relational data in synthetic DNA and also demonstrated how to perform [data processing](https://en.wikipedia.org/wiki/Data_processing) operations (similar to [SQL](https://en.wikipedia.org/wiki/SQL)) directly on the DNA as chemical processes.[[31]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-31)[[32]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-32)

In June 2019, scientists reported that all 16 GB of [Wikipedia](https://en.wikipedia.org/wiki/Wikipedia) have been encoded into [synthetic DNA](https://en.wikipedia.org/wiki/Synthetic_DNA).[[4]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-CNET-20190629-4) In 2021, CATALOG reported that they had developed a custom DNA writer capable of writing data at 18 Mbps into DNA.[[5]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-biorxiv.org-5)

The first article describing data storage on native DNA sequences via enzymatic nicking was published in April 2020. In the paper, scientists demonstrate a new method of recording information in DNA backbone which enables bit-wise random access and in-memory computing.[[33]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-33)

Davos Bitcoin Challenge[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=6)]

On January 21, 2015, [Nick Goldman](https://en.wikipedia.org/wiki/Nick_Goldman) from the [European Bioinformatics Institute](https://en.wikipedia.org/wiki/European_Bioinformatics_Institute) (EBI), one of the original authors of the 2013 [*Nature*](https://en.wikipedia.org/wiki/Nature_(journal)) paper,[[20]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-Goldman_2013-20) announced the Davos Bitcoin Challenge at the [World Economic Forum](https://en.wikipedia.org/wiki/World_Economic_Forum) annual meeting in Davos.[[34]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-34)[[35]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-:0-35) During his presentation, DNA tubes were handed out to the audience, with the message that each tube contained the private key of exactly one [bitcoin](https://en.wikipedia.org/wiki/Bitcoin), all coded in DNA. The first one to [sequence](https://en.wikipedia.org/wiki/DNA_sequencing) and decode the DNA could claim the bitcoin and win the challenge. The challenge was set for three years and would close if nobody claimed the prize before January 21, 2018.[[35]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-:0-35)

Almost three years later on January 19, 2018, the EBI announced that a Belgian PhD student, Sander Wuyts, of the [University of Antwerp](https://en.wikipedia.org/wiki/University_of_Antwerp) and [Vrije Universiteit Brussel](https://en.wikipedia.org/wiki/Vrije_Universiteit_Brussel), was the first one to complete the challenge.[[36]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-36)[[37]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-37) Next to the instructions on how to claim the bitcoin (stored as a plain text and [PDF file](https://en.wikipedia.org/wiki/PDF)), the logo of the EBI, the logo of the company that printed the DNA (CustomArray), and a sketch of [James Joyce](https://en.wikipedia.org/wiki/James_Joyce) were retrieved from the DNA.[[38]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-38)

The Lunar Library[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=7)]

The Lunar Library, launched on the [Beresheet Lander](https://en.wikipedia.org/wiki/Beresheet_lander" \o "Beresheet lander) by the [Arch Mission Foundation](https://en.wikipedia.org/wiki/Arch_Mission_Foundation), carries information encoded in DNA, which includes 20 famous books and 10,000 images. This was one of the optimal choices of storage, as DNA can last an immense period of time. The Arch Mission Foundation suggests that it can still be read after billions of years.[[39]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-39)

DNA of Things[[edit](https://en.wikipedia.org/w/index.php?title=DNA_digital_data_storage&action=edit&section=8)]

The concept of the DNA of Things (DoT) was introduced in 2019 by a team of researchers from Israel and Switzerland, including [Yaniv Erlich](https://en.wikipedia.org/wiki/Yaniv_Erlich) and Robert Grass.[[40]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-40)[[41]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-41)[[42]](https://en.wikipedia.org/wiki/DNA_digital_data_storage#cite_note-42) DoT encodes digital data into DNA molecules, which are then embedded into objects. This gives the ability to create objects that carry their own blueprint, similar to biological organisms. In contrast to [Internet of things](https://en.wikipedia.org/wiki/Internet_of_things), which is a system of interrelated computing devices, DoT creates objects which are independent storage objects, completely [off-grid](https://en.wikipedia.org/wiki/Off-the-grid).

As a proof of concept for DoT, the researcher 3D-printed a [Stanford bunny](https://en.wikipedia.org/wiki/Stanford_bunny) which contains its blueprint in the plastic filament used for printing. By clipping off a tiny bit of the ear of the bunny, they were able to read out the blueprint, multiply it and produce a next generation of bunnies. In addition, the ability of DoT to serve for [steganographic](https://en.wikipedia.org/wiki/Steganography) purposes was shown by producing non-distinguishable lenses which contain a [YouTube](https://en.wikipedia.org/wiki/YouTube) video integrated into the material.