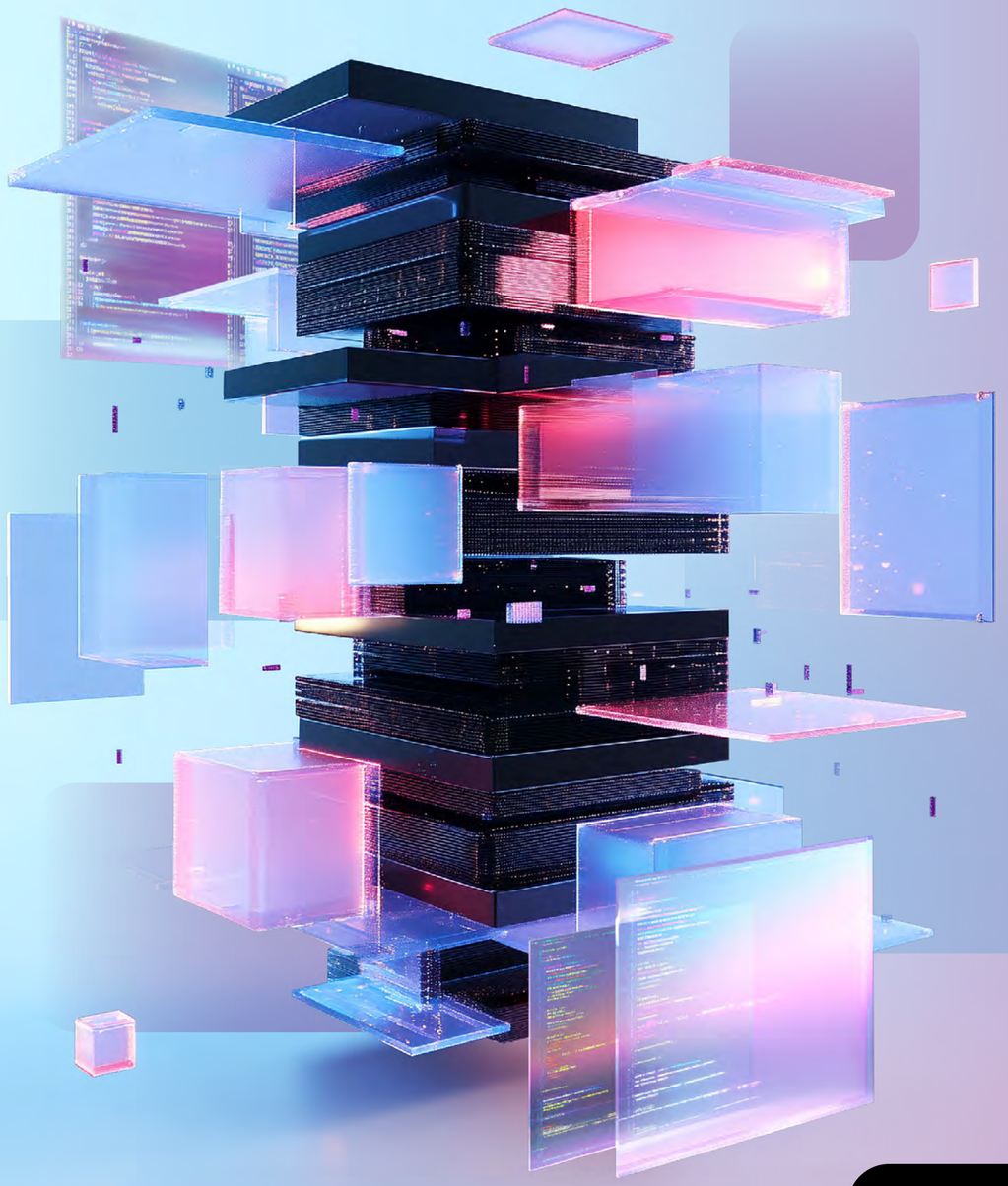
Data Management

Authors
Mark van de Sanden (GÉANT, formerly SURF), **Jan-Ru Muller** (Hogeschool van Amsterdam), **René van Horik** (DANS), **Lolke Boonstra** (TU Delft)

1. **Increased application of FAIR principles to enable digital ecosystems in Europe**
2. **Standardisation of data space architectures for secure and trusted data sharing**
3. **Growing relevance of TRUST principles for data repositories to secure data**
4. **Growing significance of augmented data management**
5. **Emergence of DNA-based data storage to preserve data for a very long time**



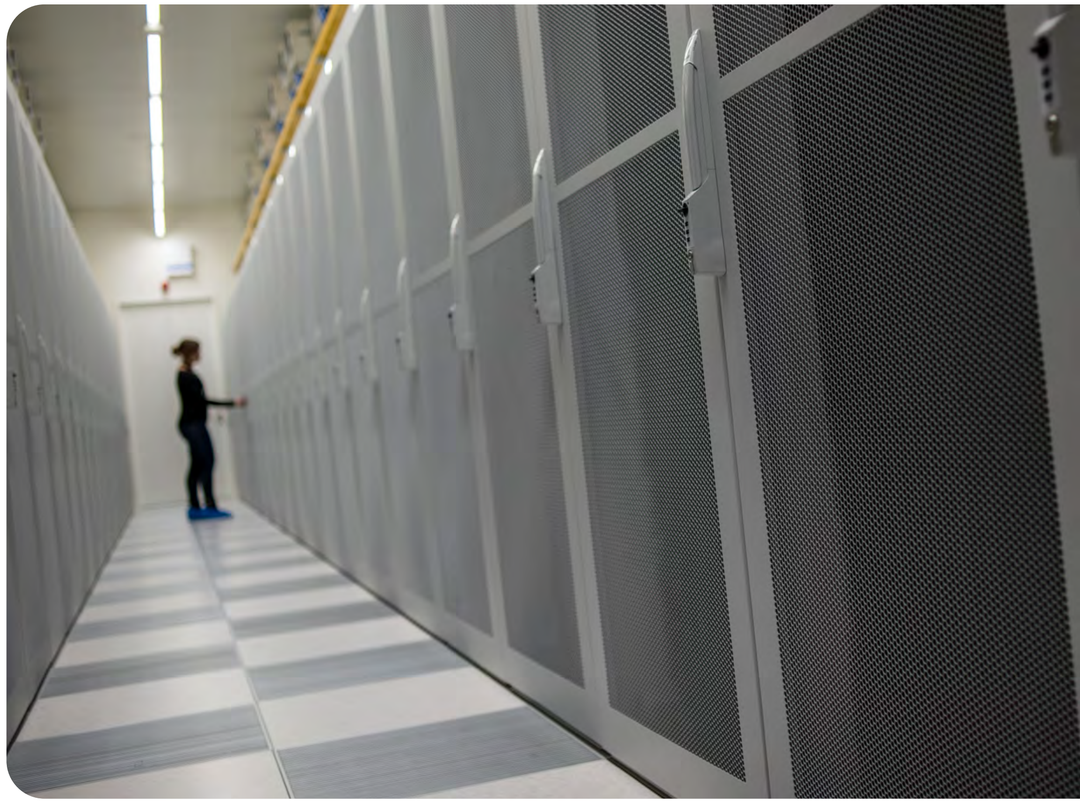
Introduction

Today's growth in data is enormous, and this growth is coming from a vast amount of sources such as industrial IoT devices (internet of things), medical imaging systems, synthetic data generators for AI model training, and big science instruments like CERN's Large Hadron Collider (particle physics) and the Vera C. Rubin Observatory (in astronomy). Current scientific experiments can already generate tens of terabytes of data on a daily basis, while future ones will push the scale to hundreds of terabytes per day.

While the current data explosion is nothing new, managing these increasing data volumes with current technologies poses challenges which require new approaches and technologies. As an example, storage media like tape and hard disk drives are reaching their physical limits regarding data densities.

Simultaneously, the variety of data sets is expanding due to various types of data. Innovative approaches and technologies are necessary, not only for the proper management of vast amounts of data, but also to combine data originating from different domains such as scientific disciplines, industries, and societal domains. This serves to enhance the capability for those managing data to uncover the key insights contained within data sets.

Next to the amount of data, the data complexity presents certain challenges. For example, on the multiple roles of organisations – like research organisations – regarding the processing of large data sets. Not only as a producer of data, not only as an user of data, but also as an actor that combines, enriches, co-creates, and aggregates large



data sets for and with a variety of other actors. Data management principles and tools help to unlock the value in data. Data management covers the systematic process of handling data throughout its lifecycle: collecting, organising, analysing, sharing, and preserving data while ensuring its integrity, accessibility, and security. AI has shown great promise already in this area, where AI-driven automation can minimise manual effort. Beyond current standard data storage solutions, there is a growing demand for data and content-aware solutions for data management as well as for offering new data insights.

Recently, it became clearer how important data management and data preservation are. Looking at developments on data sovereignty, data ownership and security, and open science. These developments are decisive for the way researchers and research organisations cooperate internationally. An example is the recent activity of the research community to preserve large climate data sets stored in the US by saving them on EU-based servers to keep the data freely available for the international climate research community. Besides this data repatriation in the scientific community, national governments and organisations in the EU are also aware of taking stronger measures to secure data ownership. For example, regarding the usage of cloud services by relocating data from the (big tech) servers in the US to servers in Europe.

New data management practices and technologies are very much on the horizon and are being formulated to tackle the current and future data challenges.

Contributors
Tom de Greef (TU Eindhoven), **Hans Tonissen** (Haskoning), **Sara Veldhoen** (Beeld en Geluid), **Rana Klein** (Beeld en Geluid), **Damiaan Zwietering** (IBM), **Laila Fettah** (IBM), **Matthijs Punter** (TNO), **Niels Bolding** (Health-RI), **Lucas van der Meer** (ODISSEI), **Marie Buesink** (Koninklijke nationale bibliotheek), **Paolo Manghi** (OpenAIRE)

TREND #1

Increased application of FAIR principles to enable digital ecosystems in Europe

Public Values

- Autonomy
- Justice — Integrity I Transparency | Sustainability
- Humanity

Maturity WATCH **PLAN** ACT

Drivers

Community dynamics & social cohesion; Globalisation; Digital transformation

The FAIR principles are guidelines on datasets being findable, accessible, interoperable, and reusable. Increased application of FAIR principles may lead to digital ecosystems, a web of FAIR data and services, where digital resources are not only shared but also semantically linked, automatically interpreted, and reliably reused across domains. Beyond data, a key evolution is the application of FAIR principles to all research outputs, including supporting frameworks such as software, computational workflows, and scientific models. This convergence is facilitated by FAIR Digital Objects (FDOs), which include persistent identifiers and rich metadata, alongside Knowledge Graphs that structure the semantic relationships between these FDOs. The shift to applying FAIR beyond just ‘data only’, to include digital research objects, enables automation, reproducibility, and information discovery, while also fostering cross-domain innovation on a larger scale.

SIGNALS

Initiatives & organisations

GO FAIR Initiative (go-fair.org)

EOSC (European Open Science Cloud) (eosc.eu)

FDO Forum (fairdo.org)

OpenAIRE (explore.openaire.eu)

> ***"The web of FAIR data and services is a keystone vision for the future of digital ecosystems."***
>
> **- Paolo Manghi**, OpenAIRE, CTO

Implementations

Aridhia FAIR Data Services - provides researchers with tools for dataset discovery, classification, and metadata browsing (aridhia.com)

Fairdata.fi - provides data storage and discovery services to support FAIR principles and ensure long-term preservation (fairdata.fi)

FAIRsharing.org - provides for a curated, informative resource on data and metadata standards, inter-related to databases and data policies (fairsharing.org)

Open Research Knowledge Graph (ORKG) - scholarly communication exploiting the possibilities of digitisation (orkg.org)

Zenodo (zenodo.org)

FAIRimpact project (fair-impact.eu)

Supporting literature

Toward the Open Science model: publish your raw diffraction data (pubs.aip.org)

Analysis on open data as a foundation for data-driven research (link.springer.com)

Leiden Declaration on FAIR Digital Objects (fdo2022.org)

IMPACT

Education

- Knowledge Graphs (KGs) are being integrated into Learning Management Systems (LMSs) such as Brightspace and Canvas to model relationships between learning objectives, content, and competencies.
- Fostering links between learner profiles, learning progress, and course requirements, KGs enable personalised learning through adaptive learning paths.

Research

- The web of FAIR data and services significantly impacts research by accelerating scientific discovery and promoting reusability.
- FAIR Digital Objects (FDOs), enriched with fine-grained metadata and versioning, enhance the reproducibility, traceability, and reuse of research data, software, and workflows.

Operations

- Knowledge Graphs (KGs) can enhance operational systems by linking data from Learning Management Systems (LMSs) and Student Information Systems (SIS).
- The integration of KGs enables institutions to map the student journey from pre-admission to graduation and subsequent alumni engagement.

TREND #2

Standardisation of data space architectures for secure and trusted data sharing

Public Values

Autonomy	Privacy | Independence
Justice	Transparency | Integrity | Accountability | Equity
Humanity	Safety

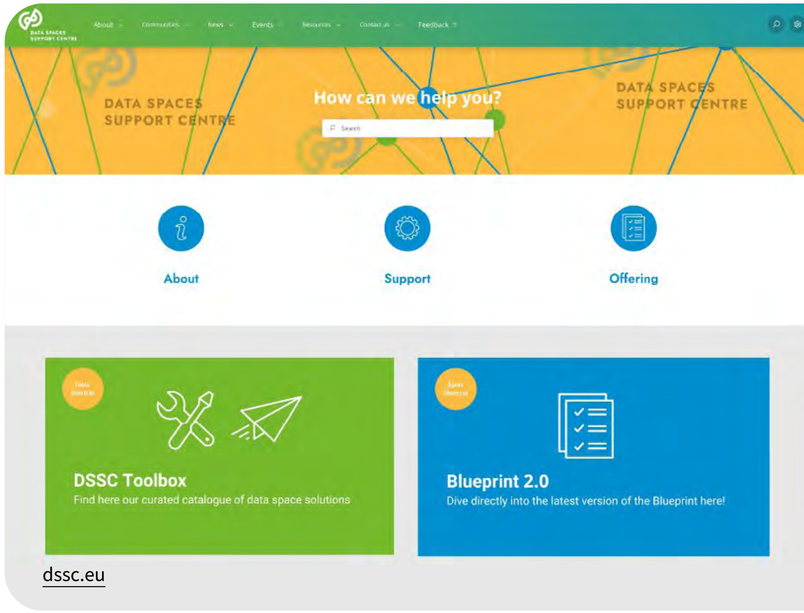
Maturity WATCH PLAN **ACT**

Drivers

Compliance & regulation; Cybersecurity & trust; Value of knowledge & skills

Data spaces allow participants to share, trade, and collaborate on data assets in a manner that is compliant with the participants' needs and regulations. These spaces could unleash the enormous potential of data-driven innovation. However, large-scale data sharing is hampered by concerns about trust and the lack of control mechanisms for sharing secure and trusted data. Therefore, standardisation of data space architectures that better support secure and trusted exchanges has received notable attention. The need for structured data spaces has grown, and considering the geopolitical situation, there are significant efforts by the European Commission to stimulate the development of common European Data Spaces.

In addition, the maturity heavily varies per domain. The maturity of a metadata standard within the domain enhances the data space's maturity. Also, the legal basis used within the domain can be very different from domain to domain.



dssc.eu

SIGNALS

European and national level Data Spaces being established

European Commission strongly supports the development of the common European Data Spaces (digital-strategy.ec.europa.eu)

Future Mobility (marketplace.future-mobility-alliance.org)

European Health Data Space (EHDS) common framework for the use and exchange of electronic health data across the EU (european-health-data-space.com)

> ***"In the early days of electricity everything was invented. You didn't come to determine what you could do with it. Now it is standardised, and you can look at where a plug can be connected. This now applies to the standardization of Data Spaces."***
>
> - **Matthijs Punter**, TNO, Data Spaces Support Centre

Reports

European Commission update on the status of the common European Data Spaces (digital-strategy.ec.europa.eu)

Critical success factors for Data Space deployment (tno.nl)

Challenges of the clean energy transition and meeting the ambitious targets of the European Green Deal (energy.ec.europa.eu)

Reference architectures being developed to support Data Spaces

Data Space Support Centre publishes version 2 of the Data Spaces Blueprint (dssc.eu)

Draft functional and technical specifications Simpl architecture published (simpl-programme.ec.europa.eu)

Npuls uses the HOSA domain architecture for education and flexibility as a cornerstone (surf.nl)

IMPACT

Education

Data spaces based on advanced blueprints enable secure, EU-compliant data sharing among educational institutions. They support AI-driven learning, ensure data sovereignty, and reduce dependence on non-European platforms.

Research

Data spaces organised by the latest blueprints in research enable secure, faster data access and facilitate cross-border sharing. They enhance scientific collaboration, accelerate research reproducibility, and drive innovation while ensuring ethical and legal data use.

Operations

By enabling secure data exchange amongst members of a data space, operations can be streamlined.
The standardised architectures enhance agility, transparency, and alignment with national and European regulations and digital transformation goals.

TREND #3

Growing relevance of TRUST principles for data repositories to secure data

Public Values

Autonomy	Privacy
Justice	Transparency | Sustainability | Accountability | Integrity
Humanity	Safety

Maturity WATCH **PLAN** ACT

Drivers

Value of knowlegde & skills; Community dynamics & social cohesion; Cybersecurity & trust; Service-oriented & value-based economies; Compliance & regulation

Data repositories are integral in digital ecosystems, facilitating long-term access to data. Therefore, TRUST principles (Transparency, Responsibility, User focus, Sustainability and Technology) are gaining more relevance recently. These principles ensure repositories are transparent in their operations, handle data responsibly and reliably, use user-focused approaches, use resources sustainably, and deploy technology to secure data management.

The TRUST principles complement the FAIR principles and enhance them by making the data infrastructure more reliable and long-term sustainable. In essence, FAIR is about the data, while TRUST is about the repository that manages and preserves the data. So this combination ensures that research data remains a reliable, accessible, and valuable resource for science and society, both now and in the future.

SIGNALS

Transparency

Open source infrastructure and open governance models

Open Science NL drives the transition to open science in the Netherlands, focusing on research data management (nwo.nl)

Open science is central to European research policy, emphasising immediate, unrestricted access to research outputs in shared repositories (research-and-innovation.ec.europa.eu)

Responsibility

Accountability compliance checks

Compliance assessment of the TRUST principles is developed by initiatives related to EOSC (faircore4eosc.eu)

Robust digital preservation strategies and a network of trustworthy repositories ensure data authenticity, integrity, and reliability (EDEN and FIDELIS projects) (eden-fidelis.eu)

Sustainability

Managing human and natural resources responsibly, e.g. by reducing energy consumption (using “green” data centres)

The “Green IT maturity model” helps organisations assess their responsibility for the environmental impact of IT and convert intentions into actionable practices (surf.nl)

Digital technology accounts for 5-9% of global electricity use, making energy efficiency vital. The EU’s green cloud initiative aims to promote energy-efficient cloud computing (digital-strategy.ec.europa.eu)

User Focus

Fine-grained data control and consent management
(as this is increasingly asked for by the user-community)

ODISSEI is a secure environment in the Dutch social science research infrastructure that facilitates easy access, sharing, and processing of sensitive data (odissei-data.nl)

The SIESTA project develops EU-level tools and methodologies for sharing sensitive data, aiming to provide researchers access to confidential information while ensuring privacy and usability (eosc.eu)

Technology

Semantic web standards and APIs enhance interoperability by providing a common language and framework for data exchange, enabling systems to understand and interact with each other more effectively.

Researchers, publications, data, funders, etc. are connected by semantic web technologies to form a “Global Open Science Graph” (graph.openaire.eu)

IMPACT

Education

The TRUST principles enable data repositories to better facilitate the education sector with the management of data in a more effective manner. They foster collaboration and shared learning through better data sharing and reuse.

Research

Researchers can (re)use and share research data more effectively when data repositories are based on the TRUST framework.

Operations

The TRUST principles can form the basis of the operational activities of next-level data repositories that consider data management dimensions like ethics, user needs, and sustainability. Such an approach will serve to enhance institutional reputation with robust and inclusive data management.

TREND #4

Growing significance of augmented data management

Public Values

Autonomy	Privacy
Justice	Transparency | Equity | Accountability | Integrity
Humanity	

Maturity WATCH PLAN **ACT**

Drivers

Automation & AI; Connectivity & interaction; Digital transformation

Data volumes are growing rapidly, increasingly generated through automated means, and AI is now accelerating this ongoing trend. Additionally, data sources such as scientific instruments, large sensor networks, and the IoT (Internet of Things) devices in general are becoming more prevalent. Managing these data volumes, extracting information, generating insights, and preserving data value is a challenging task.

Augmented data management, a form of AI-based automation, is evolving and radically reducing the manual tasks of data management teams, such as building data orchestration pipelines, assessing data quality, and running repetitive data integration workflows.

To properly harness the growing use of AI for data management, data quality is essential. Therefore, AI is recognised as an important pillar for more content-aware and data quality-aware data management solutions.

SIGNALS

Augmented data management initially recognised as trend by Gartner and Deloitte 5 years ago

Top 10 Trends in Data and Analytics for 2020 (gartner.com)

Tech Trend 2021: Machine data revolution - feeding the machine (deloitte.com)

IBM launches new content-aware storage solutions (ibm.com)

Data Direct Networks introduces the Data Intelligence Platform (ddn.com)

Augmented data management becomes mature

Development of a maturity model for AI – augmented data management (essay.utwente.nl)

Trusted European media data space (TEMS) (beeldengeluid.nl)

AI is transforming research and education

Contributing to the web of FAIR data and the uptake of AI (eosc.eu)

How AI is revolutionizing education (weforum.org)

Nplus: Ethical and effective use of AI and data (npuls.nl)

IMPACT

Education

AI-enabled data management can bring huge advantages to education, developing more personalised learning content, supporting teachers in the assessment process and supporting education by automating administrative tasks.

Research

Next to how AI is transforming how research is being done, augmented data management can automate many repetitive data management tasks in enriching metadata, quality checking, and developing and enabling more content-aware data management solutions for researchers while preserving the value of the data.

Operations

Augmented data management only works effectively on quality data. To prevent augmented data management systems from being trained with low-quality data, data managers need to be semantically skilled (on metadata, semantic vocabularies, and semantic thesauri) to ensure that high-quality data is used.

TREND #5

Emergence of DNA-based data storage to preserve data for a very long time

Public Values

Autonomy	Privacy
Justice	Sustainability | Integrity | Accountability
Humanity	

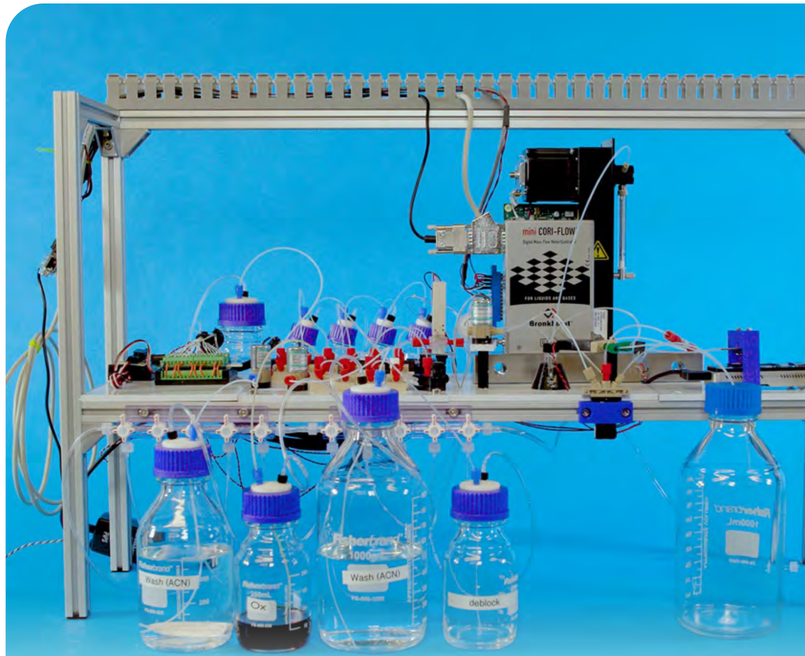
Maturity WATCH PLAN ACT

Drivers

Biotechnology; Automation & AI; Climate change & global warming; Energy supply & demand; Clean water demand; Biodiversity; Raw material scarcity

Today's storage solutions (like hard disk drives and tapes) have scaled extensively over the years; however, these types of storage media are reaching the physical limits of data storage densities. A new and promising alternative is using DNA for storing binary data in synthesised strands of DNA.

This storage solution – which will take years of R&D to be operational - could have a transformational impact on data storage infrastructures, offering potential advantages in data density, encrypted data storage, data durability, long-term data retention, and sustainability. In the meantime, R&D is focused on e.g. workable access speeds the industry needs to develop suitable standards for DNA data storage to facilitate its future deployment.



Microsoft and University of Washington demonstrate first fully automated DNA data storage. Image Microsoft (news.microsoft.com)

SIGNALS

Data synthesis companies enter the market

Twist Bioscience is a public biotechnology company established in 2013 based that manufactures synthetic DNA and DNA products for customers (twistbioscience.com)

CATALOG has created the world's first commercially viable device for DNA storage and computation (catalogdna.com)

First specifications released

The DNA Data Storage Alliance released the first two specifications: Sector Zero and Sector One (dnastoragealliance.org)

> ***"Data storage capacity of DNA per gram is around 200 million gigabytes, which is millions of times higher than magnetic tape storage densities."***
>
> **- Tom de Greef**, TU/e

Research and industry leaders are organising themselves

The DNA Data Storage Alliance established in October 2020 (dnastoragealliance.org)

The DNA Data Storage Alliance joined the Storage Networking Industry Association (SNIA) as a Technology Affiliate in 2022 (snia.org)

Institutes are experimenting with DNA Data Storage

Researchers from the University of Washington and Microsoft have demonstrated the first fully automated system to store and retrieve data in manufactured DNA (washington.edu)

Beeld en Geluid stores iconic fragments of EK'88 in DNA (nieuws.beeldengeluid.nl)

Sound & Vision (Beeld & geluid) stores iconic fragment of European Championship '88 in DNA (nieuws.beeldengeluid.nl)

IMPACT

Education

Although DNA-based data storage has no direct impact on education yet, it could become a new educational topic, or a new way for students to experiment with data storage.

Research

In research fields where data needs to be preserved for extended periods, or potentially forever due to data importance (for example, National Archive data or endangered languages), future DNA data storage solutions seem to offer long-term possibilities for preservation.

Operations

- DNA data storage could introduce an infrastructural change in how organisations will store, manage, and preserve data in the long term. This would require new skills and industry standards.
- New companies could emerge specialised in data-oriented DNA synthesis and sequencing.

More info about Data Management? Visit surf.nl

SURF Utrecht
Hoog Overborch Office Building (Hoog Catharijne)
Moreelsepark 48
3511 EP Utrecht
+31 88 787 30 00

SURF Amsterdam
Science Park 140
1098 XG Amsterdam
+31 88 787 30 00

futuring@surf.nl
www.surf.nl/en