European Innovation Council

PATHFINDER CHALLENGE
DNA-based digital data storage

CHALLENGE GUIDE
Version 14/06/2022

EIC Work Programme reference: HORIZON-EIC-2022-PATHFINDERCHALLENGES-01-05
Call deadline date: 19/10/2022 17.00 CET
Programme Managers: Roumen BORISSOV (acting)
Challenge page: https://eic.ec.europa.eu/eic-funding-opportunities/calls-proposals/eic-pathfinder-challenge-dna-based-digital-data-storage_en

The EIC will hold an online Info Session on this Pathfinder Challenge call on 05/07/2022. Participation in the meeting, although encouraged, is optional and is not required for the submission of an application. Information about how to access the Info Session and on additional dissemination events can be found at EIC Pathfinder Challenges Applicants' Day (europa.eu) **and** EIC Pathfinder (europa.eu)**.**

Table of Contents

1. About this document

The Challenge Guide serves as guidance and background for the common understanding, participation rules and obligations for the EIC beneficiaries that are involved in the Challenge Portfolio. Contractual Obligations are further detailed in the EIC work programme https://eic.ec.europa.eu/eic-work-programme-2022_en and collected in the Pathfinder Challenge guidance on contractual issues, available on the Challenge page.

The Challenge Guide is a guidance document accompanying a Pathfinder Challenge call for proposals to provide applicants with additional technical information to underpin the objectives and to provide further information about how portfolio considerations will be taken into account in the evaluation of proposals.

The Challenge Guide is prepared by and under the responsibility of the relevant EIC Programme Manager (information about the EIC Programme Managers is available on the EIC Website https://eic.ec.europa.eu/eic-communities/eic-programme-managers_en). It further details the call by complementing notably the Scope, Specific Objectives and/or Specific Conditions set out in the EIC Work Programme. In no case does the Challenge Guide contradict or supplant the Work Programme text.

2 Background concerning the scope and objectives of the Challenge

This section provides additional information on the relevant scientific and technological domains pertaining to scope and objectives of the Challenges that applicants may wish to take into account. This section should be read as background to the Challenge call in the EIC Work Programme text (attached as Annex). Proposals to this Challenge are expected to explain how they relate to and intend to go beyond the state of the art, and how they interpret and contribute to the objectives of the Challenge.

2.1 Background on DNA Digital Storage (DDS)

Data is the lifeblood of the Information Age, and the amount of data that is created, captured, and stored is growing quickly. As the world becomes increasingly digitized, the number of data sources is rising dramatically. In addition to "standard" online activities, the Internet of Things, smart cities, personal devices, autonomous vehicles, and environmental sensors all contribute to a rapid rise in the volume of data generation. Moreover, fields such as climatology, astrophysics and healthcare also generate vast amounts of data that must be retained and mined for scientific knowledge. There also exist significant potential commercial benefits for businesses engaging in long-term data storage and analytics, in terms of insights into consumer trends, and the development of new products and services.

In addition to the simple problem of more data being generated and retained, the rise in cloud-based storage solutions (with inbuilt redundancy) means that data is replicated more often nowadays, and many legislations now require the long-term retention of personal and commercial data, sometimes over several decades.

In 2020, the world generated approximately 64 zettabytes of data (a zettabyte is 10^{21} bytes, or a trillion gigabytes). By 2025, some forecasts predict global annual data generation levels of 180 zettabytes, following a compound annual growth rate of 23%[1]. Moreover, reasonable estimates suggest that, by 2025, there will exist a significant (and growing) gap between data generation and global storage capacity. That is, existing data storage technologies will be able to store less than half of the data that is being produced. A significant proportion of data that is generated is not yet stored beyond the short-term, but some longer-term projections forecast that overall demand for digital storage will exceed supply by up to three orders of magnitude by 2040.

2.1.1 Background challenges

Existing storage media are the culmination of significant scientific and technological innovations, and they are expected to form the backbone of our digital infrastructure for the foreseeable future. However, they still present significant challenges in terms of long-term and zettabyte-scale storage:

Scalability

Despite recent technological advancements, the growth in storage density for magnetic media (such as tape and hard disc drives) is slowing down. Flash storage plays a key role in short- to medium-term storage capacity, but this too is beginning to face scaling and cost roadblocks.

[1] IDC, 2021. Worldwide Global DataSphere Forecast, 2021–2025: The World Keeps Creating More Data — Now, What Do We Do with It All?, IDC Doc #US46410421.

Moreover, data scalability factors do not simply refer to storage capacity; they also refer to the number of different datasets to be handled or related to one another, the frequency with which datasets are updated and/or queried, the complexity of updates and/or queries, and the physical location, distribution, and replication of the stored datasets.

Sustainability
The Information and Communication Technology (ICT) sector (including data centres) generates up to 2% of global CO2 emissions, and data centres have the fastest growing carbon footprint across this sector[2]. It is crucial, therefore, that we investigate ways in which to reduce the energetic requirements imposed by data storage (and, by extension, its cost in terms of both financial outlay and the environment). Moreover, existing storage methods (using traditional electronic substrates) require the mining of large amounts of gold, copper, and aluminium, as well as "rare earth" materials that are difficult to process and which generate significant toxic pollution. In addition, the disposal of redundant equipment is complex and expensive, and often generates additional pollution (especially when waste is burned).

Integrity and reliability
Existing media can offer data retention periods that last on the order of decades (especially for magnetic media), but this comes at a cost in terms of proper environmental control, periodic data integrity checks, and other physical overheads. However, magnetic media are vulnerable to failure and degradation over time, and chip-based storage devices tend to have a maximum lifetime of around ten years before having to be replaced.

2.1.2 DNA as an alternative storage medium

We seek, therefore, an additional "layer" of data storage technologies that can complement existing infrastructure, offer possible solutions for long-term issues of capacity and sustainability, and support the development of novel ways in which to capture, process and propagate diverse forms of data.
One promising avenue of research focusses on the use of DNA as an alternative digital storage medium. Although the idea of using this molecule for engineered data storage dates to the 1960s, it has only recently (in the past two decades) become a practical reality[3,4]. Molecules of DNA (deoxyribonucleic acid) encode the genetic information of every living organism, and it offers several possible advantages as an artificial data storage medium:

Storage density

[2] Avgerinou, M., Bertoldi, P. and Castellazzi, L., 2017. Trends in data centre energy consumption under the European Code of Conduct for Data Centre Energy Efficiency. *Energies*, 10(10), 1470.
[3] Church, G.M., Gao, Y. and Kosuri, S., 2012. Next-generation digital information storage in DNA. *Science*, 337(6102), pp. 1628-1628.
[4] Lim, C.K., Nirantar, S., Yew, W.S. and Poh, C.L., 2021. Novel modalities in DNA data storage. *Trends in Biotechnology*, 39(10), pp.990-1003.

The raw information density of DNA has been calculated[5] as more than one exabyte (1000 petabytes) per cubic millimetre. For context, the 64 zettabytes of data generated globally in 2020 could theoretically be stored in four tablespoons of DNA[6]. Although traditional storage devices are moving beyond planar layouts to 3D structures, the inherent volumetric storage of DNA may also confer significant potential benefits.

Robustness
DNA is remarkably stable[7], especially when stored in carefully controlled conditions (maintaining pH, temperature, humidity, etc.) Even in sub-optimal conditions, DNA can remain intact and readable for hundreds or even thousands of years. This means that DNA will not suffer the same level of degradation as magnetic (or even chip-based) media, especially when carefully encapsulated for long-term ("cold") storage applications.
Universality
The genetic code (the mapping between nucleotide sequences and amino acids that make up proteins) is shared by all living organisms, meaning that it is one of the best-studied encoding schemes in existence. This universal code stands in contrast to data encoding standards, many of which have become obsolete over time, leading to difficulty in decoding information. We may be confident, though, that technologies to read DNA sequences will exist in the future; moreover, developments in the life sciences will continue to further refine and optimise these methods, which may then find application in DNA storage (of course, this field may also, in turn, drive developments in DNA synthesis, storage, and read-out methods).

Resource requirements
Once data is properly encoded in DNA, it requires very little resource for its long-term storage[8]. Although most of the energy used by traditional data centres is expended on reading, writing and copying data (as opposed to simply storing it), and these operations on DNA still generally require significant resources, if progress in DNA-based technologies continues at the current rate then we can expect DNA-based media to become competitive relatively quickly, especially for long-term storage.

2.1.3 Dynamic DNA storage

Although much recent attention has focussed on the use of DNA as a long-term storage medium, we should not overlook its potential as a dynamic medium. Although many of the proposed applications of DNA storage currently assume a relatively static database, we might imagine a future need to be able to handle changing data, as well as the ability to implement

[5] Bornholt, J., Lopez, R., Carmean, D.M., Ceze, L., Seelig, G. and Strauss, K., 2016. A DNA-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 637-649.
[6] 64 zettabytes=64,000 exabytes. At 1mm^3 per exabyte, 64,000 cubic millimetres=4.32 US tablespoons.
[7] Matange, K., Tuck, J.M. and Keung, A.J., 2021. DNA stability: a central design consideration for DNA data storage systems. *Nature Communications*, 12(1), pp. 1-9.
[8] Goldman, N., Bertone, P., Chen, S., Dessimoz, C., LeProust, E.M., Sipos, B. and Birney, E., 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*, 494(7435), pp.77-80.

database operations within (as opposed to simply "on") the molecular storage medium[9]. A recent illustration of this used preferential binding for similarity searching on a DNA-based image database (that is, the physical properties of the DNA molecules were used to implement pattern matching)[10].

2.1.4 Living cells as possible storage media

Most demonstrations of DNA storage use in vitro molecules. However, the natural environment of DNA is, of course, the living cell, and evolution has produced a diverse toolbox of biological processes for copying, proof-reading and even propagating DNA sequences. In addition, recent developments such as CRISPR offer unprecedented opportunities for direct manipulation of microbial genomes for the purposes of data storage[11]. A seminal application of this stored the frames of a short movie inside bacterial genomes[12], and a possible major advantage of this approach is the ease of replication of data stored inside the cell, using its inherent "machinery". Perhaps more significantly, in vivo data storage also offers the possibility of living cells acting as "smart" devices for sensing and recording[13].

3 Portfolio considerations for the evaluation of applications to the Challenge

This section describes how portfolio considerations will be taken into account in the second stage of the evaluation of applications. In the first stage, all applications will be evaluated individually by external experts and scored against the evaluation criteria set out in the Work Programme. All applications that pass the defined thresholds against the criteria will be included in the second stage of the evaluation. At the second stage, all above threshold applications will be considered collectively by an evaluation panel chaired by a relevant Programme Manager. At this stage, the Evaluation Committee will consider which applications to recommend for funding in terms of a coherent portfolio of projects that can interact, reinforce or compete with each other to increase the overall impact.

Categories

This EIC DDS pathfinder call and the associated portfolio of projects has the aim of consolidating and extending Europe's already strong position in DNA digital storage. EIC DDS thematic portfolio activities will nurture funded projects, as well as encouraging networking within the innovation ecosystem. The call will fund a portfolio of breakthrough and complementary projects on molecular storage. In the second evaluation step, the evaluation committee, chaired by the Programme Manager, will build a consistent portfolio of projects to achieve specific strategic objectives.

---

[9] Lin, K.N., Volkel, K., Tuck, J.M. and Keung, A.J., 2020. Dynamic and scalable DNA-based information storage. *Nature Communications*, 11(1), pp.1-12.

[10] Bee, C., Chen, Y.J., Queen, M., Ward, D., Liu, X., Organick, L., Seelig, G., Strauss, K. and Ceze, L., 2021. Molecular-level similarity search brings computing to DNA data storage. *Nature Communications*, 12(1), pp.1-9.

[11] Yim, S.S., McBee, R.M., Song, A.M., Huang, Y., Sheth, R.U. and Wang, H.H., 2021. Robust direct digital-to-biological data storage in living cells. *Nature Chemical Biology*, 17(3), pp.246-253.

[12] Shipman, S.L., Nivala, J., Macklis, J.D. and Church, G.M., 2017. CRISPR–Cas encoding of a digital movie into the genomes of a population of living bacteria. *Nature*, 547(7663), pp.345-349.

[13] Sheth, R.U. and Wang, H.H., 2018. DNA-based memory devices for recording cellular events. *Nature Reviews Genetics*, 19(11), pp.718-732.

During the assessment process, the evaluation committee will assign proposals to categories, which define the overall architecture of the portfolio. All proposals will be mapped against these categories, and the evaluation committee will look for complementarities in the following categories:

Storage duration: The committee will seek to facilitate a set of projects that operate across the spectrum of time, from long-term, archival, "cold" storage, to medium-term, "working" storage and short-term "dynamic" storage.

Substrate: Whilst acknowledging that the bulk of the portfolio may be comprised of projects that operate on synthetic DNA *in vitro*, the committee will encourage the investigation of alternative substrates such as non-natural polymers or living cells.

Portfolio considerations

The following considerations will be applied to all proposals in the ranking list, which are above the thresholds:

1) If present, at least one proposal will be added to the portfolio dealing with each of short-, mid, and long-term storage, since each of these time-frames presents its own set of challenges (in terms of the stability of the molecules used, and the technologies required to manipulate them) and opportunities (in terms of possible applications).
2) Portfolio construction will encourage a degree of diversity in the underlying substrate (DNA, other synthetic polymers, living cells, etc.), whilst acknowledging the current emphasis on synthetic DNA.
3) The committee will encourage proposals that demonstrate end-to-end integration and interoperability, if appropriate to the category.
4) Within each category, the committee will seek, where possible, diversity in terms of potential applications (so, for example, within the "long-term" window, they may look for applications that require different levels of fidelity of data storage).

4 Implementation of the Challenge portfolio

Once selected, projects will be expected and obliged to work collectively during the implementation of their projects under the guidance of an EIC Programme Manager. This section summarises some of the key aspects of this pro-active management which applicants should take into account in preparing their proposals.

Grant negotiations

Applicants may be requested to make amendments to their proposed project in order to take into account the portfolio objectives and enhance the portfolio. Such changes may include: an additional work package to undertake common/ joint activities (workshops, data exchanges, joint research, etc) with other projects in the portfolio; adjustments to the timings of some

activities and deliverables in order to synchronise better with the implementation timings of other projects; specific target changes to improve complementarity/ comparability with activities and results from other projects. All such changes will be discussed during the grant preparation stage with the aim of reaching a consensus between all projects on the adjustments needed.

Challenge portfolio roadmap

Following the selection of proposals to be funded under the Challenge, the Programme Manager will work together with the selected projects to develop a common roadmap for the Challenge. This roadmap will integrate the activities and milestones of the individual projects into a shared set of objectives and cross-project activities. The roadmap serves as a common basis for implementing the projects - including possible adjustments, reorientations or additional support to projects - and can be updated in light of emerging results or difficulties during the implementation. The objectives can be revised, for instance based on projects' unexpected achievements, new technology trends, external inputs (other projects, new calls…).

In particular, the Challenge roadmap will include activities on the transition to innovation and commercialisation, and to stimulate business opportunities. These activities may be supported and reinforced during the implementation with additional funding and expertise through pro-active management.

Pro-active management of projects

Projects in the portfolio may be offered additional support, either individually or collectively, in order to reinforce portfolio activities or explore the transition to innovation. Such additional support includes:

- Booster grants of up to €50k (see Annex 6 of the EIC Work Programme)
- Access to additional EIC Business Acceleration Services (see [link])
- Access to the Fast Track to the EIC Accelerator, the decision for which would follow a project review (see Annex 4 of the EIC Work Programme)
- Access to the EIC Market Place, once operational, to connect with innovators, investors and other selected partners
- Interactions with relevant projects and initiatives outside the portfolio, including other EU funding initiatives as well as those supported by national, regional, or other international bodies.

ANNEX Extract of EIC work programme

II.2.5 EIC Pathfinder Challenge: DNA-based digital data storage

Introduction and scope

Current technologies for digital data storage are hitting sustainability limits in terms of energy consumption and their use of rare and toxic materials. Moreover, data integrity when using those technologies is limited in time, which complicates archival data-storage. DNA or certain classes of synthetic DNA alternatives provide an alternative that promises information densities that are several orders of magnitude higher than classical memories, and stability for millennia rather than years. Moreover, DNA-based data storage can profit from the growing range of DNA research, tools and techniques from the life sciences, while potentially also adding to it (e.g., for in-vivo data collection).

Proof of concept for DNA data archiving in vitro (i.e. not in living cells) is now well established. Several studies have shown that such archiving can support selective and scalable access to data, as well as error-free storage and retrieval of information. However, technical challenges remain to make this process economically viable for a broad spectrum of uses (beyond so-called 'cold data') and data types. These relate to improving the cost, speed and efficiency of technologies for reading, and especially writing and editing, DNA or other information-storing biopolymers.

Large corporates and governments are starting to show an interest and some smaller companies offer solutions for specific archival applications. Europe has academic and commercial potential in this area. The time is right to pull together a European R&I ecosystem on DNA-based digital data storage.

This EIC Pathfinder Challenge is to explore scalable and reliable high-throughput approaches for using DNA as a general data-storage medium. Solutions would thus need to address the read/write/edit operations of digital data in synthetic DNA, capturing the expected advantages of high density and stability/longevity of this form of data storage. The use of DNA sequences as chassis for non-standard forms of information coding, or of other polymeric substrates and related coding/decoding techniques are also in scope, provided they entail at least similar benefits than state-of-the-art DNA approaches. Proposed techniques should deliver qualitative advances in key parameters such as throughput, DNA-length (well above a few hundred meters), reliability (coupling efficiency), speed and cost. Beyond the usual storage applications, there is also scope for radically different scenarios for such a technology, for instance for data-processing, in-vivo sensing or fingerprinting. 41

Applications submitted to this Challenge, must pay particular attention to the relevant biosafety and ethical issues.

Specific objectives

The following specific objectives for this Challenge have been defined:

- new approaches for coding, decoding, modification or computational use of digital data in synthetic DNA or other sequence-controllable polymers with quantitative targets (theoretical and technological);
- Proof-of-Concept of technical feasibility with indications of at least state of the art benefits and major operational characteristics (e.g., extreme densities, longevity, stability) and going well beyond for some of them (e.g., speed, cost, accuracy);
- end-to-end scenarios of use, be it for data storage (archival, but also shorter term storage) or other purposes (like sensing, cryptography or computation) that exploit the benefits of the technology.

Expected outcomes and impacts

Proposals should contribute to achieving one or several of the following:

- a range of new techniques with clear benefits and steps towards widening scope of applicability of DNA-based data storage;
- broader range of scenarios and uses for DNA-based data technologies;
- emergence and anchoring of a European innovation eco-system on DNA-based data technologies and applications, including through involvement of relevant partners and end-users;
- contribution to standardisation in the field and benchmarks to gauche progress.

Specific conditions

Proposals for this Challenge can be submitted by single applicants or by consortia, as dictated by the activities to be performed.