h e g

Haute école de gestion
Genève

Archives Cantonales Vaudoises

OAIS compliant digital archiving in DNA

Travail de Bachelor réalisé par :

Dina ANDRIAMAHADY

Sous la direction de :

Basma MAKHLOUF SHABOU, professeure HES

Genève, 14 juillet 2021

Information documentaire

Haute École de Gestion de Genève (HEG-GE)

Hes·so Genève
Haute Ecole Spécialisée de Suisse occidentale

Déclaration

Ce Travail de Bachelor est réalisé dans le cadre de l'examen final de la Haute école de gestion de Genève, en vue de l'obtention du titre Bachelor of Sciences HES-SO en information documentaire.

L'étudiant atteste que son travail a été vérifié par un logiciel de détection de plagiat.

L'étudiant accepte, le cas échéant, la clause de confidentialité. L'utilisation des conclusions et recommandations formulées dans le Travail de Bachelor, sans préjuger de leur valeur, n'engage ni la responsabilité de l'auteur, ni celle du conseiller au Travail de Bachelor, du juré et de la HEG.

« J'atteste avoir réalisé seule le présent travail, sans avoir utilisé des sources autres que celles citées dans la bibliographie. »

Fait à Genève, le 14 juillet 2021

Dina ANDRIAMAHADY

Remerciements

Je voudrais témoigner toute ma reconnaissance à toutes les personnes qui ont contribué de loin ou de près à la réalisation de ce Travail de Bachelor.

Ma conseillère pédagogique, Madame Basma Makhlouf Shabou, pour le soutien qu'elle a apporté au projet ainsi que son encadrement et sa disponibilité tout au long du travail.

Mon mandant, Monsieur Jan Krause-Bilvin pour son enthousiasme et son accompagnement mais surtout pour le temps qu'il a consacré au projet.

Un grand merci à toutes les personnes que j'ai rencontrées et qui ont enrichi le contenu du présent travail à travers différents échanges.

Merci à tous ceux qui ont accepté de relire ce travail en particulier Aurèle Nicolet, Améthyste Bovay, Floriane Muller et Linda Meiser. Leurs conseils ont été particulièrement appréciés.

Toute ma gratitude et ma reconnaissance vont à ma famille pour leur soutien et leurs encouragements pendant ces trois ans et surtout leur patience et leur amour inconditionnel.

Abstract

Time and space have been among the main concerns for archivists for a very long time. In other words, how to keep information accessible and understandable for the longest time and how to organize the ever-growing amount of data. Average storage media longevity is short from an archival point of view and similarly, technological progress leads to media obsolescence. Furthermore, media decay over time as well. The sheer volume of information that professionals have to handle yearly remains a current and pressing matter.

DNA is one of those “words” that everyone has heard but that most have only a nebulous grasp. DNA stands for deoxyribonucleic acid; it is a biomolecule that holds the genetic information of all living beings. DNA stores and transfers the information necessary for the development and maintenance of living organisms to the next generation. As such, it possesses astounding characteristics, honed by nature’s genius for billions of years, making it appealing as a digital data storage medium.

Particularly, its longevity and its remarkable information density appear to settle the above preoccupations. However, is it a viable solution in an archival system?

To operate a DNA-based storage in an archival system a prerequisite would be a conformance to the Open Archival Information System (OAIS), the reference model in authority in the world of libraries, archives and museums.

This implementation will also necessitate a definition of technical characteristics that will allow an effective and efficient implementation of this media to a long-term preservation system. It is essential to define the confines of this endeavor: a structural and tangible framework within which DNA technology can be transposed into.

The present work aims to establish which requirements are necessary for the system to meet the OAIS compliance and which effective technological solutions to use for an optimal fulfillment of the model through study of literature and discussion with preservation practitioners and DNA applications experts.

Digital archiving in DNA is definitely feasible in accordance with the reference model and existing DNA-storage protocols. However, currently used DNA technology impedes its viability: costs and latency remain particularly high preventing an immediate application to an archival system. Different methods and protocols for information writing, access, storage and reading are being investigated to optimize the use of DNA as storage medium. In time, these technological advances will allow a competitive usage of DNA as the ultimate tool for information preservation.

Keywords: *Digital preservation, deoxyribonucleic acid (DNA), Open Archival Information System (OAIS)*

Table of contents

List of figures

Glossary

AIP Edition

*“An AIP whose Content Information or Preservation Description Information has been subject to an upgrade or improvement which was not required for preservation. An AIP Edition is not considered (CCSDS 2019)d to be the result of a Migration.”(*CCSDS 2019)

AIP Version

*“An AIP resulting from changing the Content Information or Preservation Description Information of a source AIP, in order to preserve the information. An AIP Version is considered to be the result of a Migration.”(*CCSDS 2019)

Archival Information Package (AIP)

“An Information Package, consisting of the Content Information and the associated Preservation Description Information which in preserved within an OAIS.” (CCSDS 2019)

Archive

“An organization that intends to preserve information for access and use by a Designated Community.” (CCSDS 2019)

Archives

"n., [records] The whole of the documents made and received by a juridical or physical person or organization in the conduct of affairs, and preserved. Syn.: fonds.

n., [place] A place where records selected for permanent preservation are kept.

n., [institution] An agency or institution responsible for the preservation and communication of records selected for permanent preservation." (InterPARES [n.d.])

Base pairs (bases)

“Pair of complementary nucleotides in DNA.” (WIB [n.d.])

Content Information

“A set of information that is the original target of preservation. It is an Information Object composed of its Content Data Object and its Representation Information” (CCSDS 2019)

CRUD

“ CRUD (Create, Read, Update, Delete) is an acronym for ways one can operate on stored data. It is a mnemonic for the four basic functions of persistent storage. CRUD typically refers to operations performed in a database or datastore, but it can also apply to higher level functions of an application such as soft deletes where data is not actually deleted but marked as deleted via a status.” (MDN [n.d.])

Data Object

“Either a Physical Object or a Digital Object.” (CCSDS 2019)

Descriptive Information

"The set of information, consisting primarily of Package Descriptions, which is provided to Data Management to support the finding, ordering, and retrieving of OAIS information holdings by Consumers." (CCSDS 2019)

Designed Community

"An identified group of potential Consumers who should be able to understand a particular set of information in ways exemplified by the Preservation Objectives. The Designated Community may be composed of multiple user communities. A designated Community is defined by the Archive and this definition may change over time." (CCSDS 2019)

Dissemination Information Package (DIP)

"An Information Package, derived from one or more AIPs, and sent by Archives to the Consumer in response to a request to the OAIS." (CCSDS 2019)

DNA (Deoxyribonucleic Acid)

"[A] complex chemical found in the nucleus and mitochondria of a cell. It provides the genetic instructions needed for an organism to develop, survive and reproduce." (WIB [n.d.])

DNA microarray

"Also known as a DNA chip, a DNA microarray comprises a collection of microscopic DNA spots printed on to a solid surface that is used to measure the expression levels of large numbers of genes simultaneously." (WIB [n.d.])

DNA polymerase

"DNA polymerase is a type of enzyme that can be found in all living organisms. There are many types of DNA polymerase. Some help replicate DNA when a cell divides and others help in the day-to-day repair and maintenance of DNA." (WIB [n.d.])

DNA sequencing

"A biochemical method used to determine the exact order of the four building blocks, nucleotide bases that make up a piece of DNA." (WIB [n.d.])

Double helix

"Term used to describe the spiral configuration of DNA." (WIB [n.d.])

Double-stranded

"A molecule that consists of two bound strands, each of which complements the other. DNA is usually double-stranded." (WIB [s.d.])

Fixity Information

*"The information which documents the mechanisms that ensure that the Content Data Object has not been altered in an undocumented manner."(*CCSDS 2019)

Information Object

"A Data Object together with its Representation Information." (CCSDS 2019)

Information Package

"A logical container composed of optional Information Object(s). Associated with this Information Package is Packaging Information used to delimit and identify the Information Object and the optional Package Description information used to facilitate searches for the Information Object." (CCSDS 2019)

Nucleic acid

"Long molecule made up of smaller molecules called nucleotides which are chemically linked together in a chain. Nucleotides are instrumental in transferring genetic information from one generation to another. There are two types of nucleic acids: are deoxyribonucleic acid (DNA) and ribonucleic acid (RNA)." (WIB [n.d.])

Nucleotides

"Nucleotides are molecules present in all cells of the body. They serve many purposes, including acting as cellular messengers between the outside and the inside of a cell's nucleus, storing energy and as physiological mediators. Nucleotides are also necessary to the construction of the nucleic acids DNA and RNA. DNA is made up of four base nucleotides: adenine (A), cytosine (C), guanine (G) and thymine (T). RNA is made up of A, G, and C, plus uracil." (WIB [n.d.])

Oligonucleotides

"Commonly made in the laboratory, an oligonucleotide is a short sequence of DNA (usually 2-50 bases). Oligonucleotides are important in artificial gene synthesis, polymerase chain reactions, DNA sequencing, library construction and can be used as molecular probes." (WIB [n.d.])

Packaging Information

"The information that describes how the components of an Information Package are logically or physically bound together and how to identify and extract the components. " (CCSDS 2019)

Polymerase chain reaction (PCR)

"A technique that is used to copy a specific DNA sequence. The technique provides the means to make one billion exact copies of an original target DNA within a couple of hours." (WIB [n.d.])

Preservation Description Information (PDI)

"The information, which along with Representation Information, is necessary for adequate preservation of the Content Data Object and which can be categorized as Provenance Information, Context Information, Reference Information, Fixity Information, and Access Rights Information" (CCSDS 2019)

Primer

"Template strand of DNA used to generate a new double-strand of DNA." (WIB [n.d.])

Record

"n., A document made or received in the course of a practical activity as an instrument or a by-product of such activity, and set aside for action or reference. Syn.: archival document." (InterPARES [n.d.])

Refreshment

“A Digital Migration where the effect is to replace a media instance with a copy that is sufficiently exact that all Archival Storage hardware and software continues to run as before.” (CCSDS 2019)

Repackaging

“A Digital Migration in which there is an alteration in the Packaging Information of the AIP.” (CCSDS 2019)

Replication

“A Digital Migration where there is no change to the Packaging Information, the Content Information, and the PDI. The bits used to represent these Information Objects are preserved in the transfer to the same or new media-type instance.” (CCSDS 2019)

Representation Information

“The information that maps a Data Object into more meaningful concepts so that the Data Object may be understood in ways exemplified by Preservation Objectives.” (CCSDS 2019)

Submission Information Package (SIP)

“An Information Package that is delivered by the Producer to the OAIS for use in the construction or update of one or more AIPs and/or the associated Descriptive Information” (CCSDS 2019)

List of abbreviations and acronyms

The following table gives the significance of the various abbreviations and acronyms used throughout this work. Not all of them are official denominations but they allowed us to shorten and facilitate the designation of entities.

Abbreviation	Meaning
ACV	Archives cantonales vaudoises
AIP	Archival Information Package
BnF	Bibliothèque nationale de France
bp	Base pair
CCSDS	Consultative Committee for Space Data Systems
CERN	Conseil européen pour la recherche nucléaire
CRC	Cyclic Redundancy Check
CRUD	Create, Read, Update and Delete
DIP	Dissemination Information Package
DLCM	Data Life-Cycle Management
DNA	Deoxyribonucleic Acid
DRUID	Digital Resource Unique Identifier
ETHZ	Eidgenössische Technische Hochschule Zürich
FEDORA	Flexible Extensible Digital Object Repository Architecture
HEG-GE	Haute École de Gestion de Genève
HES-SO	Haute École spécialisée de Suisse occidentale
ISO	International Organization for Standardization
LOCKSS	Lot Of Copies Keep Stuff Safe

NDIIPP	National Digital Information Infrastructure and Preservation Program
NGS	Next Generation Sequencing
nt	Nucleotide
OAIS	Open Archival Information System
OASIS	Organization for the Advancement of Structured Information Standards
OCFL	Oxford Common File Layout
ONT	Oxford Nanopore Technologies
PANDORA	Preserving and Accessing Networked Documentary Resources of Australia
PARADISEC	Pacific And Regional Archive for Digital Sources in Endangered Cultures
PCR	Polymerase Chain Reaction
PDI	Preservation Description Information
SIP	Submission Information Package
SPAR	Système de Préservation et d'Archivage Réparti
TdT	Terminal deoxynucleotidyl Transferase
TRAC	The Trustworthy Repositories Audit & Certification: Criteria and Checklist
UCSD	University California San Diego
UKDA	UK Data Archive
UNIGE	Université de Genève
UTS	University of Technology Sydney

1. Introduction

Digital data is being generated continuously in this Information Age and even though most of it will be discarded, a good proportion has to be stored. The trends lean obviously toward a production rate increase that will apparently intensify in the years to come. More than ever, digital preservation is confronted with the need to handle this staggering amount of data in a sustainable way. Sustainability must be considered since at present time, storage solutions are extremely energy-intensive. Another concern points at long-term persistence: digital preservation is meant to last a very long time but storage media degradation and obsolescence cause data losses and pose challenges to preservation making migrations mandatory processes.

Deoxyribonucleic acid or DNA carries the genetic information that allows the perpetuation of life from one generation to another. DNA stores and transfers genetic information and is known to be a stable and perennial molecule. Another one of its assets is the density of information it can contain. Furthermore, DNA being at the foundation of life makes obsolescence unaffordable so it will always be paramount to harness DNA pertaining technology.

Those features make DNA a particularly attractive option as an information storage medium. Since the 1980s, research groups have been looking into the possibility of exploiting these DNA characteristics for the storage of digital data. This solution seems to bring a good number of answers to the issues that affect digital preservation namely lifespan, robustness, storage scalability, energy consumption and low risk of obsolescence.

To instate DNA as a digital information storage medium, it is necessary to assess how DNA technology can be integrated into a digital preservation system. Open Archival Information System (OAIS) reference model is the ISO standard (ISO 14721) that is used worldwide to define digital preservation concepts and requirements. DNA as a storage medium will have to be consistent with OAIS specifications to effectively adhere to an archival preservation system.

1.1 Mandate

Jan Krause-Bilvin, archivist for the Canton of Vaud state archives, the Archives cantonales vaudoises (ACV), initiated the present work. It is an auxiliary study conducted in conjunction with the development of a new document governance system within the ACV. The two projects are independent.

This research project is an exploratory work that aims at identifying new storage and preservation possibilities in the field of digital archiving.

1.2 Purposes

This work is set on analyzing the feasibility of modeling a new preservation technique. This new practice must be consistent with the OAIS reference model. This work is intent on identifying the OAIS specifications directly related to the implementation of this new storage medium then assess how DNA technology can meet and fulfill the requirements.

In addition, we will work at identifying the means necessary for this implementation.

2. Methodology

First and foremost, the present work is an exploratory study bringing together two fields that have never been confronted. Information storage in DNA is not a new research subject. However, placing it in an archival framework such as the OAIS reference model is a first.

OAIS reference model is a generic and agnostic framework used to define the features required in a long-term digital repository. It does not offer implemental building blocks that can be used instantly. Therefore, it was mandatory to set basic elements to create a tangible system in the beginning.

The ACV were very interested in investigating and learning more about the potential of Oxford Common File Layout (OCFL), a standard on how objects could be organized and stored in repositories, to support the file system. To define OAIS and OCFL specifications, a literature review was conducted, followed by insights from resource persons during expert interviews. Only then was it possible to see what DNA technology would need to cover in order to be transposed to an archival storing system.

The first part of this study consists in exploring those specifications and determine the criteria DNA have to meet in order to conform to these standards.

For the second part of the project, a systematic literature review was necessary to assess the extent of what had already been done in the DNA information-storing field of study. An ensuing expert interview allowed an ascertaining of how those findings could be applied to the identified file system architecture.

Before undertaking any action, a meeting was set up with our academic advisor, Dr. Basma Makhlouf Shabou, in order to define the scope of the work and academic expectations as well as identify informational resources. As supervisor to this project, she also provided guidance on the project directions and finally ensured that the chosen approach was appropriate.

2.1 Digital preservation implementation

2.1.1 Literature search

The main sources of information we looked into were the OAIS (CCSDS 2019) and the OCFL (Hankinson *et al.* 2020) specifications in order to determine then describe the requirements the system needed to fulfill, particularly in the Archival Storage entity.

Our academic advisor provided us with additional research leads as well as the texts describing the ISO: 14721 standard.

In addition, we aimed to clarify the meaning of “OAIS-compliant”. We were interested in approaches and experiences in OAIS conformance for archival institutions and repositories. To explore this topic of interest, we set up the following research questions:

- How to use the OAIS model as a framework to set up digital preservation in institutional digital repositories?
- Which institutions have documented their strategies to achieve OAIS compliance?

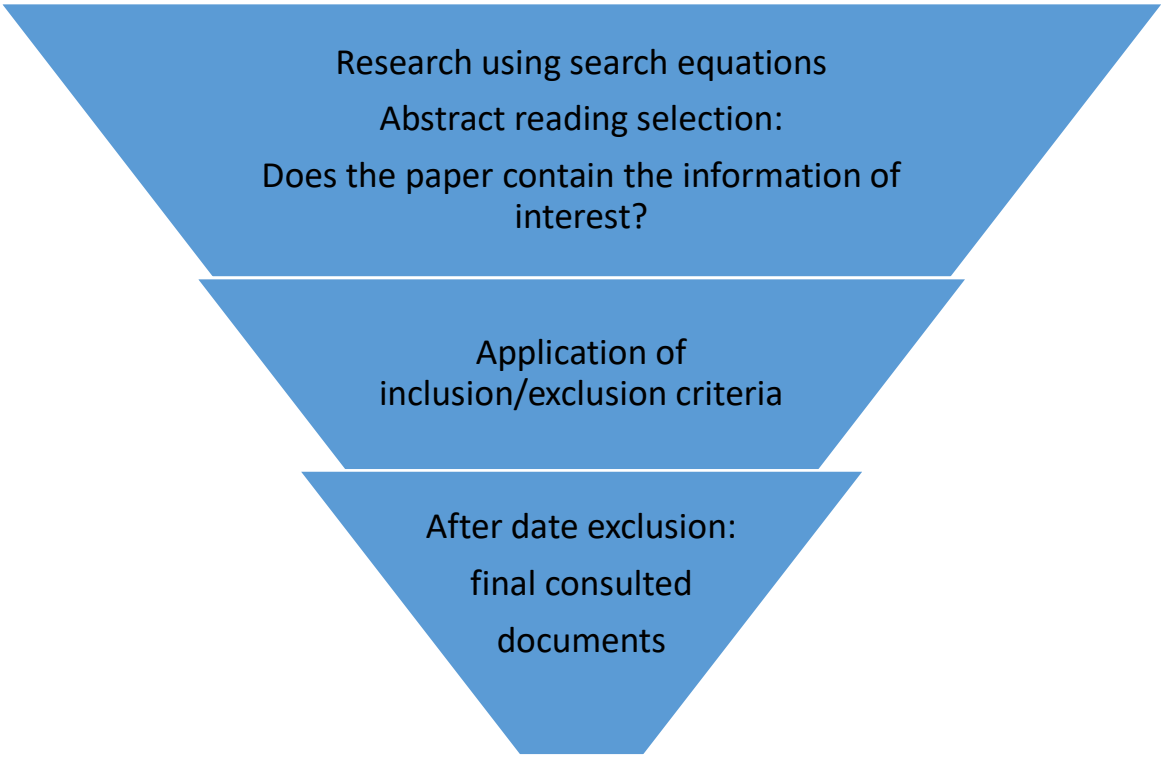
- What conceptual and applied factors are considered important for digital preservation and what requirements should a storage system meet to achieve OAIS conformance.

We looked into various resources telling about implementations of OAIS using specialized database such as LISA (Library & Information Science Abstracts) and LISTA (Library, Information Science & Technology Abstracts) and the following research equations:

- "digital preservation" AND (OAIS OR "Open Archival Information System")
- (OAIS OR "Open Archival Information System") AND implementation
- (OAIS OR "Open Archival Information System") AND complian*

We then proceeded to a selective treatment of information in order to retain the most relevant to our goals. The main criterion for inclusion and exclusion of studies was its practical aspect. Then, date was used as an exclusion factor: all papers prior to 2012 were discarded since the reference model was reviewed in 2012.

Figure 1 : Documents selection process to define "OAIS-compliance"



We decided to retain the theoretical studies when they broached the OAIS-compliance topic though, as well as seminal articles about OAIS such as the one written by Lavoie in 2004. To diversify our sources, we also consulted several archival institutions websites. They were the most informative about institutions implementation processes and OAIS-compliance evaluation.

2.1.2 Expert interview

We conducted an "Expert interview" in accordance with what is said in the book *Interviewing Experts* (2009):

“talking to experts in the exploratory phase of a project is a more efficient and concentrated method of gathering data than, for instance, participatory observation or systematic quantitative survey.” (Bogner *et al.* 2009, p. 2)

Expert interviews are a subtype of focus groups with fewer but more select participants . These interviews offer a wide range of advantages in that information is readily accessible in a less time consuming and more informal way. Moreover, experts have valuable knowledge borne of theoretical studies and empiric experiences. Finally, experts tend to be motivated persons often interested in further collaboration and exchange on the topic of interest (Van Audenhove 2011).

2.1.2.1 Instrument development

The choice of a semi-structured session was necessary because it is first to present a synthesis of what was drawn from the literature review and the reflections carried out with the client regarding the specifications and secondly, to obtain feedback from the experts and generate new ideas. A presentation was prepared beforehand to delimit and clarify the subject to direct the discussion but, more importantly, to generate new ideas and concepts so that the specifications would be as exhaustive as possible.

2.1.2.2 Resource persons

The experts were chosen because, together with their institutions, they have developed and implemented their own electronic archiving systems that comply with the ISO 14721 standard. Every one of these systems is a different but correct interpretation of the international norm. These people have a concrete and precise knowledge of the OAIS model and its requirements.

This perception of who can be an expert is consistent with the concept formulated by Meuser and Nagel (2009) describing the latter as a person who contributed to the development or the implementation of a solution, a strategy or a policy in a certain domain.

We aimed for diversity to cover the widest range of potential applications of the system. We have therefore targeted people from:

- docuteam SA, an information management enterprise working broadly with administrative records.
- Université de Genève (UNIGE) and its comprehensive research disciplines needed a data management strategy that was completed with the development of Yareta and OLOS solutions powered by the DLCM project in which the UNIGE team contributed actively.
- CERN an organism at the forefront of data management technology.

Moreover, these persons contribution and their work in the field of electronic archiving is recognized and acknowledged internationally.

Finally, they are people who expressed a genuine enthusiasm for the project and they were sincerely looking forward to seeing the evolution and the outcome of the present work.

2.1.2.3 Expert interview sessions

The proceedings took place between March and April 2021, in three virtual 90 minutes sessions. The interviews were held in French. Consequently, the interview report and presentation appended to the present work are also in French.

Each time, in addition to the experts, Mr. Jan Krause-Bilvin served as an observer during the meeting and Dina Andriamahady acted as the moderator of the interview. The observer lent support to moderator in notes taking. This led to a debriefing and sharing of the notes taken.

The encounters were not recorded since expert interviews allow a less formal setting. We were interviewing one person at a time, and thoroughly taking notes. According to Rutakumwa (2020):

> *"The comparison of the data quality between audio-recorded transcripts and interview scripts written directly after the interview indicated that they were comparable in the detail captured. ».*

In addition, Muet (2003) recommends against recording since transcription is particularly time-consuming.

2.2 DNA as a storing medium in the OAIS Archival Storage Entity

2.2.1 Systematic literature review

The use of DNA as a medium for storing information did not materialize until early 2010, even though the idea has been around since the mid-sixties. Trying to be as thorough as possible, a systematic review was conducted to examine the findings in the DNA of digital information to date. With the aim of being as exhaustive as possible, we strived to obtain all literature, both published and unpublished.

We used Web of Science as the main interface to submit our search requests. Afterwards, we used referred and related articles to identify more work pertaining to the topic. We then proceeded to write a synthesis listing the landmarks of DNA preserving digital information history.

We started with the following research equations:

- DNA AND "digital preservation"
- DNA AND "digital information"
- DNA AND "digital storage"

The selection process was less sequential than above. Once we identified a few reviews, as mentioned before, we looked into cited and citing references to identify more papers for our literature review.

2.2.2 Expert interview

2.2.2.1 Instrument development

As mentioned above, we opted for a semi-structured session. A presentation was prepared in order to direct the discussion, to present a synthesis of what was drawn from the literature review, and the reflections carried out regarding the specifications and how DNA could be used in this particular setting.

2.2.2.2 Resource persons

The main key informant person we reached out to was a titular professor at the Functional Materials Laboratory whose research group has been working on using DNA as a medium for

digital information since 2015 at ETH Zürich (ETHZ). Their research aims to make DNA an alternative device for the long-term preservation of data.

An interview was scheduled with a doctoral student from the team, currently undertaking a doctoral study on the digital applications of DNA. In 2020, a paper they published offered a step-by-step protocol on how to process data for preservation in DNA: encode it into a DNA sequence, synthesize the corresponding DNA molecule, encapsulate it in silica beads, then extract it again, sequence the DNA to extract the information, decode it then finally read it.

2.2.2.3 Expert interview sessions

The interview took place in May 2021, as a virtual 90 minutes session. An interview report and presentation were added to this work's appendix.

In addition to the expert, Mr. Jan Krause-Bilvin served as an observer while Dina Andriamahady acted as the moderator of the interview. The observer and the moderator conferred in a debriefing session after interview.

Once more, the encounters were not recorded because of the reasons stated above.

3. Open Archival Information System

The Open Archival Information System (OAIS) is a reference model used to set up a standard structure for the long-term preservation of digital information. OAIS is an Archive as in a structure, constituted of people and systems, which took upon itself to accommodate and maintain information then make it available for specific users known as Designated Community. Its “Archive” status implies its obligation to fulfill a certain number of responsibilities stipulated in the Consultative Committee for Space Data Systems (CCSDS) recommendation.

The term “Open” in the OAIS denomination should be taken cautiously. It should not be understood as in “unrestricted access” to information, but rather it was used to indicate the participatory process that generated the said recommendation.

According to Thomas, a reference model “is an information model used for supporting the construction of other models.” (2006, p. 492). As such, OAIS describes the requirements a digital preservation system should meet to fulfill its digital preservation stewardship. It is an “abstract framework” (OASIS [n.d.] cited in Schumann and Recker 2013) that sets forth the vocabulary, the concepts and the components of a digital preservation system. For this reason, it does not provide any specifications for an actual implementation.

3.1 Background

The Consultative Committee for Space Data Systems (CCSDS) developed the OAIS Reference Model in 2002. The CCSDS is an international spatial forum acting for the development of communications and data handling standards for space research. As early as the 1970s, space agencies had to contend with the processing and the storing of an astronomical amount of digital data from space experiments. In addition, they experienced firsthand the woes of digital technologies obsolescence (Huc [n.d.]).

From the 1980s, CCSDS produced a number of recommendations for space data system standards (Kummer 1987) and since 1990, through close cooperation with the International Organization for Standardization (ISO), those releases can undergo ISO review and voting in order to become formal ISO standards.

ISO then asked CCSDS to work on the elaboration of standards dedicated to the long-term archiving of spatial data. Since no widely recognized model existed, CCSDS set to create a common framework. As long-term preservation of digital artefacts was not restricted to space matters, CCSDS decided to make the model making process accessible to any interested individual or institution.

The reference model was crafted through a participative and iterative process. The first draft (for review) was released in 1997 and, in 2000, CCSDS published a draft ISO standard. After ISO reviewing and revision, the reference model became what is known as international ISO standard 14721 (Lavoie 2004).

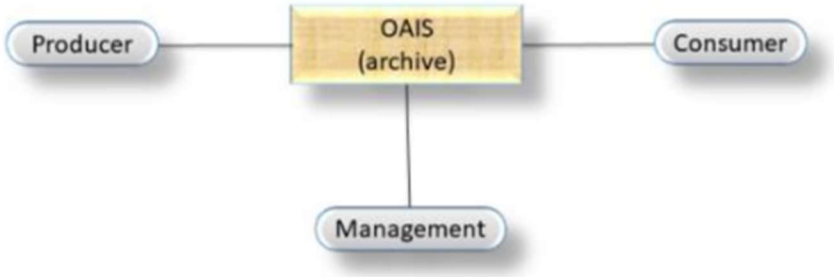
In the following chapter, we will clarify the reference model specifications with an overview of the content of the latest CCSDS draft (CCSDS 2019). This is done to identify in a practical way the concepts and requirements a storing system must observe to be in conformance with the aforementioned standard.

3.2 Open Archival Information System reference model

3.2.1 OAIS environment

The OAIS model exists within an entire ecosystem the same way an archive is an organ within an institution commissioned to cater to its community. It is called the OAIS environment and comprises four different constituents that interact through the OAIS: the **producer** of the information, its **consumer**, the **archive** (OAIS) at the center and the **management** in charge of its running.

Figure 2 : OAIS Environment diagram



(CCSDS 2019, p. 2-2)

3.2.2 OAIS Information definition

To qualify as an OAIS Information Object, the main data (Data Object) intended for upkeep must be associated with “Representation Information”, a piece of information documenting said data (how it was made, what is it made of, what media and format were used to contain it, …) in order to allow its decryption at any given time.

Figure 3 : From data to information

(CCSDS 2019, p. 2-4)

For practicality in describing the exchanges occurring between OAIS and information producers and consumers, CCSDS came up with the idea of the Information Package, a conceptual container carrying the Content Information (the object to preserve, described above as Information Object) and Preservation Description Information.

There are three types of information package within the OAIS system:

- The **Submission Information Package** (SIP), which is the incoming package from producers.

- The **Archival Information Package** (AIP), which consists of one or more SIPs taken in for preservation.
- The **Dissemination Information Package** (DIP), which is a part or an entire AIP made available following a consumer request.

3.2.3 OAIS functional entities

Six main functional entities compose an OAIS. These entities fulfill the roles encompassing the interactions between the four members of the OAIS Environment. These roles are briefly described and explained in the following section.

Figure 4 : OAIS Functional Entities

(CCSDS 2019, p. 4-1)

3.2.3.1 The Ingest Functional Entity

Ingest Function receives and accepts SIPs from producers. Once the SIP is verified, an AIP can be generated from it afterwards the AIP is transferred to the archival storage.

3.2.3.2 The Archival Storage Functional Entity

Archival Storage Entity, or Archival Storage for short, is in charge of storing, maintaining and retrieving the AIP. AIPs accepted from Ingest are added to permanent storage. Archival Storage functions include media replacement when necessary, routine error checks, reconstruction policy enforcement in case of disaster and consumers requests fulfillment.

3.2.3.3 The Data management Functional Entity

Data Management works mainly on curating Descriptive Information and maintenance of administrative data used to operate the Archive.

3.2.3.4 The Administration Functional Entity

Administration oversees the running of operations within the archival system. It is charged with negotiating submission agreements with information producers, SIPs quality auditing, customer's services and maintenance and improvement of overall Archive performances.

3.2.3.5 The Preservation Planning Functional Entity

Preservation Planning provides recommendations and preservation plans to keep the information accessible and understandable even over a very long storage period. Among its proposals are information updates, migration recommendations, periodic risk analysis reports, technology innovation.

3.2.3.6 The Access Functional Entity

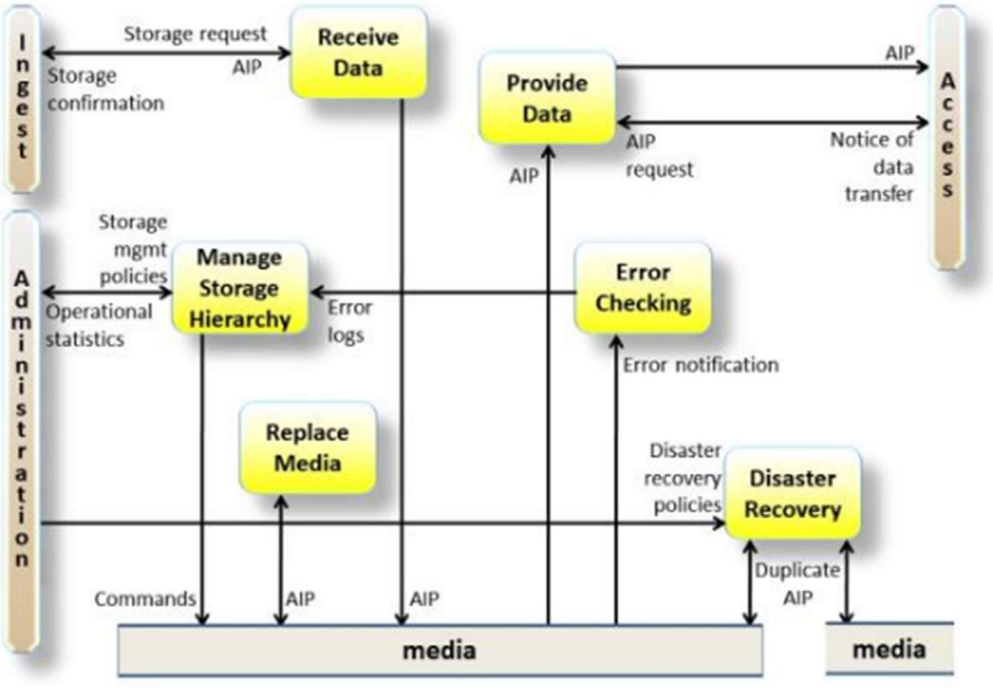
Access provides the means for consumers to identify and retrieve information from OAIS. Its role consists in communicating with consumers, enforcing access policy if necessary and delivering the required DIP as requested.

Now that we have a general understanding of how OAIS systems operate, we will focus more particularly on Archival Storage functions where lay our main interest in the present work.

3.2.4 Archival Storage Functional Entity specifications

The main goal of this work is to see how and to what extent one can implement DNA as a medium for long-term digital information preservation in an OAIS compliant system. Understanding the different functions in the Archival Storage is mandatory in order to define how DNA can play a role in this module.

Figure 5 : Archival Storage Functional Entity



(CCSDS 2019, p. 4-8)

3.2.4.1 Receive Data function

The “Receive Data” part is about accepting the SIP and defining the best media type for optimal preservation: for instance, if the information needs to be accessed often, it must be assigned

to an appropriate storage device. This function is in charge of adding new elements to Archival Storage. The AIP reception is documented by the sending of a storage confirmation message to Ingest.

3.2.4.2 Manage Storage Hierarchy function

“Manage Storage Hierarchy” is extremely important since it defines how AIPs are distributed throughout the storage hierarchy. Certain AIPs will require specific handling due to particular status (such as a need for special security measures). The function will also monitor objects access rates and system error tolerance. Error logs monitoring by the Manage function will guarantee that information is not damaged nor corrupted during transfer. It will also perform backup processes. It also has to keep Administration abreast on the Archive performances, storage capacity and general state.

3.2.4.3 Replace Media function

The “Replace Media” function ensures the system’s competency to replicate the AIPs in the long term. This reproducibility must take into account the following stipulation by CCSDS (2019, p. 4-9):

> *“Within the Replace Media function the Content Information and Preservation Description Information (PDI) must not be altered. However, the data constituting the Packaging Information may be changed as long as it continues to perform the same function and there is a straightforward implementation that does not cause information loss.”*

Most of the time, media replacement consists in a migration to different media (at different levels: Refreshment, Repackaging or Replication) in order to preserve information. Preservation plans must make provision criteria for media selection. Selection is done in consideration of media error rates, performance, cost and longevity.

3.2.4.4 Error Checking function

Routine Error Checks called Cyclic Redundancy Check (CRC) are performed and statistics are provided to Manage Storage Hierarchy to ensure system integrity in Archival Storage or during data transfer. A Fixity Information guarantees that data has not been modified during AIP retrieval and subsequent access. Additional tracing methods could be set up and error detection and correction protocols can be appended by Archive.

3.2.4.5 Disaster Recovery function

The mandatory system recovery policy is achieved by setting up duplication procedures for the whole Archive collection. It is often carried out by keeping duplicates in different locations, mainly physically distanced amenities. Even though the Disaster Recovery function carries out the effective measures, it is the Administration entity role to set up the disaster recovery policies.

3.2.4.6 Provide Data function

Here the main task consists in supplying Access with a copy of AIP on request. Upon receiving an AIP request that allows its identification, the Provide Data function produces a copy of AIP that is then sent to Access Functional Entity. A notice of data transfer mailed to Access documents the operation.

3.2.5 Archival Information Package

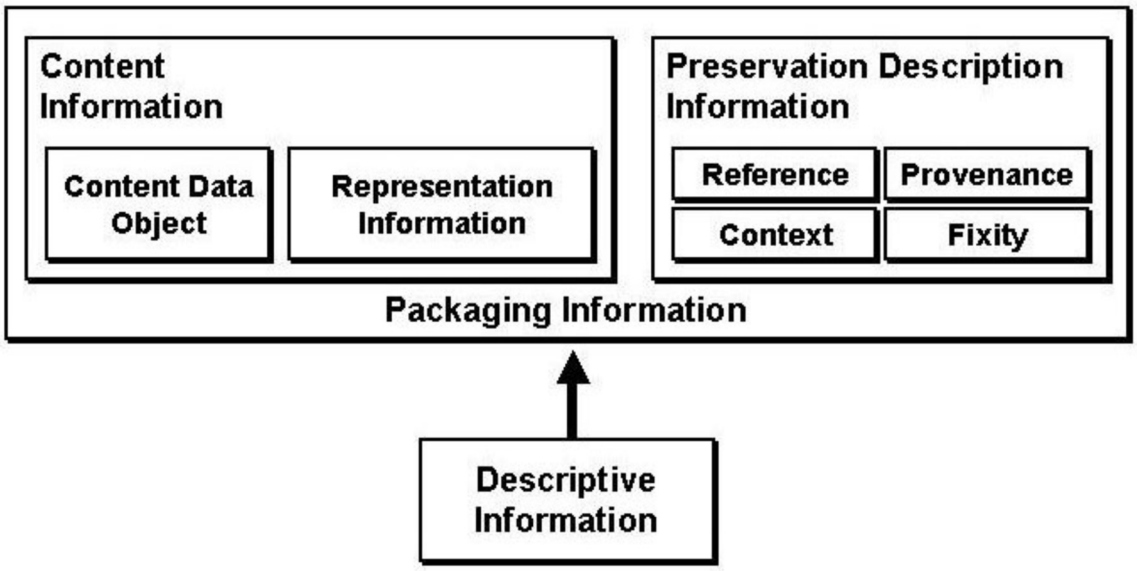
AIPs are the units used to store the information in the Archival Storage Functional Entity. We already defined the Information Package as being a "*conceptual structure for supporting Long Term Preservation of information*" (CCSDS 2019, p. 4-36).

An AIP is the more specific subtype of the Information Package that is actually stored and preserved by the Archive. It is composed of an Information Object called Content Information that is the actual artefact intended for preservation. Beside the Content Information, the AIP contains the Preservation Description Information (PDI). The latter describes the Content Data Object that is part of the Content Information.

The AIP itself has its own descriptors. Those are the Packaging Information that can be inscribed into the media holding the AIP. In adjunction, a Package Description accompanies the AIP. It serves as an equivalent to a bibliographic record aiding in identification, location and retrieval of information by consumers.

The AIP consists in information destined for retention and its complete set of metadata aiding in its preservation and access. These metadata are not always physically attached to the information they describe. However a logical link must exist between the two and they are seen as a "*logical package within the archival system*" (Lavoie 2004, p. 75).

Figure 6 : Archival Information Package diagram



(Lavoie 2004, p. 76)

3.2.6 OAIS compliance

CCSDS defines OAIS compliance as "*supporting […] its Mandatory Responsibilities and its Information Model*" (CCSDS 2019, p. 6-10). The OAIS recommended practice content devotes a paragraph to describing what an archival system should submit to be in conformance with the model. This is what it says:

"*A conforming OAIS Archive implementation shall support the model of information described in 2.2. The OAIS Reference Model does not define or require any particular method of*

implementation of these concepts. A conforming OAIS Archive shall fulfill the responsibilities listed in 3.1." (CCSDS 2019, p. 1-3)

This delimitation of OAIS compliance is carried over by Ball (2006) as well as Chandra and Gokhale (2012).

The expression "OAIS implementation" is often mistakenly used when referring to the setting up of an archival system in conformance with the standard. Schumann and Recker (2013, p. 10) aptly describe that situation and they later sustain that "*what OAIS compliance means is really a matter of interpretation*".

However, there are tools used to audit digital repositories. The Trustworthy Repositories Audit & Certification: Criteria and Checklist, shortened as TRAC, is one of them. As the name suggests, it is a checklist of all the requirements a repository must observe in order to be in conformance with OAIS. TRAC elaboration was based on OAIS principles. At the conclusion of the audit, a score between 0 and 4, where 4 means "complete conformance", is delivered rating the repository. LOCKSS was submitted to such an evaluation in 2014 and made available the audit results (LOCKSS [n.d.]) which is very informative of the procedure. TRAC successor, ISO:16363, is used to certify trustworthy repositories.

3.2.7 Systems using the OAIS Reference Model

Since its installation in 2002, many institutions have premised their digital preservation system on OAIS (Arnold, Denis *et al.* 2020; Ghadami 2020; NDIIPP 2002). They range from libraries, national archives, scientific data centers and commercial organizations to initiatives and projects. Among the most well-known organizations that adopted OAIS are the National Archives and Records Administration, the Library of Congress and the British Library. Listed below are a few other systems that have opted to conform to OAIS:

- The National Archives and UK Data Archive (UKDA)
- The National Information Infrastructure and Preservation Program (NDIIPP) set up by the Library of Congress to preserve digital heritage;
- The PANDORA project (Preserving and Accessing Networked Documentary Resources of Australia) destined to host Australian on-line publications;
- DSpace, a repository system from DuraSpace, widely used (among others: World Bank, World Health Organization, Massachusetts Institute of Technology).
- SPAR (Système de Préservation et d'Archivage Réparti) developed by La Bibliothèque nationale de France (BnF) to support its digital efforts was launched in 2010.

OAIS is currently under review since ISO and CCSDS regulations expect a revision every five years. Back during its revision in 2009, Nicholson and Dobreva were adamant that the way it was "*OAIS is not enough in itself to ensure the successful preservation of digital materials*" (Nicholson, Dobreva 2009, p. 8).

OAIS is accused of being incomplete and vague, leaving an open field of interpretation and making it more difficult to define what "OAIS compliance" means.

"In consequence, there is no guidance on what constitutes a minimum technical requirement with respect to integrating functional bit stream processing into corresponding system designs

which would help in the real-life implementations of DP [digital preservation] systems." (Nicholson, Dobreva 2009, p. 3)

The OAIS recommended practice text states on several occasions that it does not endorse any implementation tools. It is a guide that layers the principles on which a digital archiving system should be built. It offers common grounds for all archival institutions and repositories to talk about and compare their systems.

A practical structure and method on how to store data in a digital preservation system are necessary in order to actually implement a storage system module in an archival system.

3.3 Oxford Common File Layout

We established that the OAIS reference model was an abstract set of terminologies, notions and responsibilities that an organism intent on long-term digital preservation should follow. It does not give actual building blocks for effective implementation of an archival infrastructure. In this work, it was necessary to have a practical approach to how files are actually stored in the file system to be able to devise its DNA feasibility. The ACV was interested in seeing how the Oxford Common File Layout (OCFL), a recently published file layout standard, would fare.

The OCFL proposes what a college of specialists think how repositories should organize and store objects as bits on a file system (Hankinson *et al.* 2020). It is "*an open community effort defining an application-independent way of storing versioned digital objects with a focus on long term digital preservation*" (Hankinson *et al.* 2019, p. 1).

3.3.1 Background

The OCFL initiative came about during a conversation between Andrew Hankinson and Andrew Woods during the Fedora and Hydra/Samvera camp they attended in Oxford in September 2017.

In March 2018, people from Cornell University, Stanford University, DuraSpace, Oxford University and Emory University constituted an editorial group to work with the community on specifications and use cases. In September of the same year, they gathered at Oxford University to start writing the Oxford Common File Layout specification (Hankinson *et al*. 2019; 2020).

The OCFL tackles five specific requirements pertaining to issues often encountered but never addressed in a unanimous way or that are still unresolved (Hankinson *et al.* 2019; Jefferies, Bredenberg, Dappert 2019).

- **Completeness**: the system carries all the data and metadata required for repository reconstruction.

Figure 7 : OCFL logo symbolizing completeness after a disaster

(Hankinson *et al.* 2020)

- **Parsability**: readability by humans and machine ensuring there is no need to have software mediating access to stored digital object.
- **Robustness** against errors and corruptions and an ability to support migrations and error checks are major prerequisites expected from any digital preservation system.
- **Versioning** allows management and access to older and more recent versions of files and rendering object history.
- **Storage interoperability**: the ability to store content on different infrastructures and to migrate it indifferently between different systems, which provides some measure of protection against obsolescence and system failure.

Before OCFL, there were two main approaches to data layout within digital repositories: BagIt and the Moab design.

3.3.1.1 BagIt

BagIt is "*a set hierarchical file layout conventions for storage and transfer of arbitrary digital content*" (Kunze *et al*. 2018, p. 1) born of a Library of Congress (LoC) and California Digital Library (CDL) collaboration in 2008. It was named after the "enclose and deposit" method also known as "bag it and tag it".

BagIt is a simplified packaging structure that allows a secure data transmission using physical media or network transfer. Its main feature consists of a "bag" comprising the content files, the **payload**, and the associated metadata, the **tags**.

BagIt structure demands the following mandatory elements:

- "a set of required and optional tag files" (Kunze *et al.* 2018);
- "a subdirectory named 'data', called the payload directory" (Kunze *et al.* 2018);
- "a set of optional tag directories" (Kunze *et al.* 2018).

BagIt, initially designed as a file transfer mechanism tool, was later widely used as a unit of storage supporting whole archival structures.

3.3.1.2 The Moab design

Moab was created by Richard Anderson and named after the place where he lives (Utah, USA).

Figure 8 : Moab folder structure (left) - Natural Arch in Moab, Utah (right)

```
ab123cd5678
   v0001
      data
         content
            title.jpg
            intro.jpg
            page1.jpg
            page2.jpg
            page3.jpg
         metadata
            versionMetadata.xml
            descMetadata.xml
            identityMetadata.xml
      manifests
         versionInventory.xml
         signatureCatalog.xml
         versionAdditions.xml
         fileInventoryDifference.xml
         manifestInventory.xml
   v0002
      data
         content
            page2.jpg
         metadata
            versionMetadata.xml
            technicalMetadata.xml
      manifests
         versionInventory.xml
         signatureCatalog.xml
         versionAdditions.xml
         fileInventoryDifference.xml
         manifestInventory.xml
```

(Anderson 2013)

(LoggaWiggler 2008)

Together with a team from the Stanford Libraries' Digital Library Software & Services, Anderson devised a versioned, forward-delta AIP format (Anderson 2013) after using BagIt for a long time. The latter was lacking functionalities essential to digital object lifecycle handling the team needed for their reshaping of the Stanford Digital Repository (Hankinson *et al.* 2019).

The main features that characterize the Moab design are:

- Versioning through a forward-delta process (subsequently supporting deduplication);
- Parsability using human-comprehensible XML files;
- Easy random files pick up by using a unique identifier, the digital resource unique identifier (DRUID);
- Fixity completed by the use of checksum technology on every file;
- Full Object recovery (for any version).

OCFL relies heavily on the framework Moab offered, keeping its key features and addressing its flaws (Hankinson *et al.* 2019). Moab is seen as the OCFL direct parent.

3.3.2 Oxford Common File Layout specifications

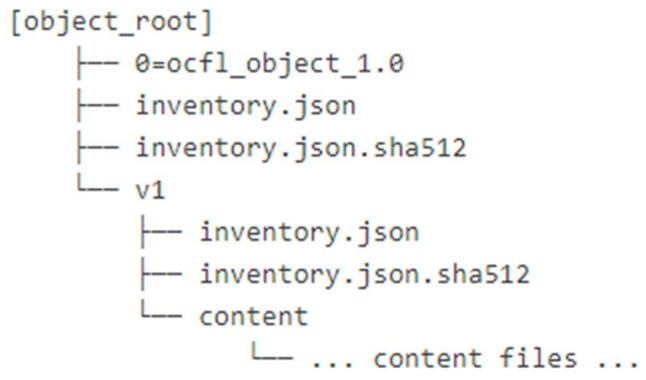
OCFL is a way of arranging files on a hierarchical file system. It is how digital repository practitioners see "the ideal layout and characteristics for a repository's persisted objects" (Hankinson *et al.* 2019, p. 1).

3.3.2.1 OCFL Objects

In our OCFL-based model, OCFL objects represent AIPs. An OCFL object consist of:

- A directory with one or numerous content files and their associated administrative information, namely an inventory listing all existing files up to this version;
- An inventory in JSON format and named **inventory.json** that defines the object's current version number and lists the object contents;
- Said inventory's digest;
- A conformance declaration as a NAMASTE file at the OCFL Object root.

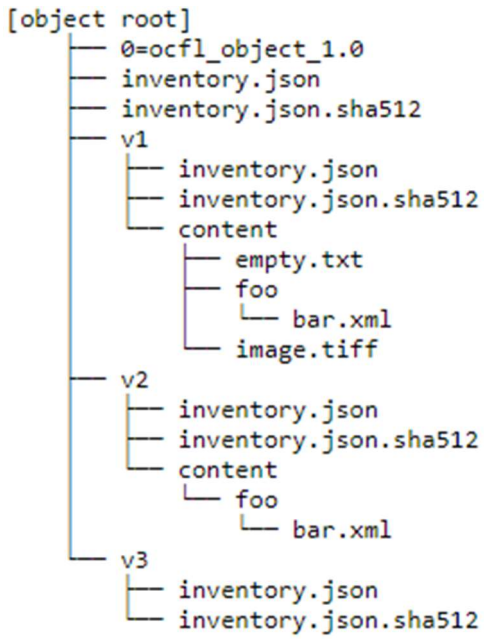
Figure 9 : Composition of a minimal OCFL Object



(Hankinson *et al.* 2020)

Versions directories are named as the version number and can be found in the Object. Every one of those folders contains an inventory and its digest to ensure information integrity. Inventories include a manifest block that lists all the digests and existing file paths. Separation of the logical file path from the existing file path allows deduplication since digests point to former versions that remain immutable.

Figure 10 : Example of an OCFL model with three version

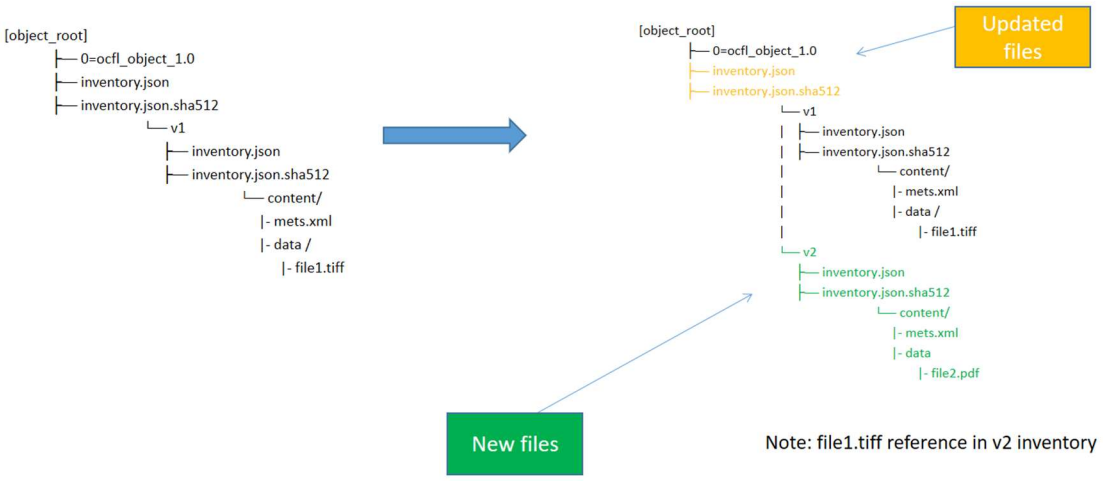


(Hankinson et al. 2020)

Following is an example of versioning and deduplication: on the left is an initial OCFL Object containing the first version (v1) of a file. On the right side is the result of versioning: a new file, named “file2.pdf” in this example, is added and contained in a v2 folder.

“file2.pdf” has its own set of metadata, an inventory and its digest. Deduplication is seen in the fact that the content of v1 is not replicated in v2. However, the inventory in v2 holds a reference pointing to the existence of version 1.

Figure 11 : OCFL versioning and deduplication example



3.3.2.2 OCFL storage hierarchy

OCFL hierarchy places the AIPs directly at the base just beneath the OCFL Storage Root. OCFL Storage Root “is the base directory of the OCFL storage layout” (Hankinson *et al*. 2020) and delimitates the storage hierarchy. It comprises a Root Conformance Declaration stating the OCFL version to which the objects comply. To deal with filesystem constraints, objects are stored in different directories and it is necessary to use a mapping system that locates the file with a unique identifier. Pairtree is a solution that supports interoperability. It uses AIP identifier, pairing characters to generate a deterministic path locating the object to its folder.

3.3.3 OCFL applications

3.3.3.1 Fedora 6

Fedora (Flexible Extensible Digital Object Repository Architecture) is an open source adaptive repository since 2003. Its latest version, Fedora 6 released on June 13th, 2021, will be integrating OCFL. Subsequently, all Fedora 6 based solutions such as Islandora 8 will integrate an OCFL layer (Jordan, Barnes 2019).

3.3.3.2 Chronopolis

Thanks to a grant by Andrew W. Mellon Foundation, the University California San Diego Library will work on a digital preservation platform part of network of distributed digital preservation systems using the OCFL standard (Luna 2019).

3.3.3.3 University of Technology Sydney

The University of Technology Sydney (UTS) set up a research data repository using the OCFL (Sefton 2019; Mike Lynch 2019).

Figure 12 : OCFL object screenshot from UTS

(Michael Lynch 2019)

Through their collaboration with PARADISEC[1] (Pacific And Regional Archive for Digital Sources in Endangered Cultures), whose application is also operating over an OCFL repository, they developed the Arkisto, a platform that makes data research available. The Arkisto is also based on OCFL standards (Sefton 2020).

[1] Proof of concept of an OCFL-based repository https://mod.paradisec.org.au/about

3.3.4 OCFL perspectives

Even though the OCFL is still a work in progress, its founding principles stem from people and institutions versed in the practice of digital preservation. As a result, it has benefitted from over twenty years of collective experience. Moreover, a strong and active community supports the OCFL editorial board. Even before the release of the beta version, many institutions have tried out the OCFL specifications, the most significant work being done by, namely, the John Hopkins University, Oxford Research Archive, University of Wisconsin-Madison, Cornell University and Stanford University. The open process is particularly advantageous to the elaboration of specifications. Relatively frequent community meetings and use cases submissions support the effort. Version 1.1 is expected by the end of the year 2021. OCFL progress is definitely worth keeping up with for any digital preservationist.

3.4 OCFL as the framework of a novel DNA-based digital archiving

3.4.1 Initial considerations

The OCFL appeared particularly appealing because of the principles on which it was built. Storage diversity and portability between different file systems and object storage allowed the flexibility we needed since at that point of the project we did not know what would be possible and what constraints we would be confronted with the use of DNA. What is more, the possibility of indifferent migrations between different systems would facilitate DNA implementation.

The built-in system of multiple checksums ensured the recording of all operations, guaranteeing information integrity in addition of supporting the error check processes required by OAIS.

The delta-forward method inherited from Moab permits simplified versioning, which is unavoidable since, at the very least, migrations require that metadata be updated.

The combination of delta-forward versioning and the inventory system was fundamental to deduplication within AIPs allowing referencing between versions and subsequently a certain relief in the archiving system with the inutility of keeping the main content in each new version.

In addition, at the moment OCFL is made to be tried and experimented. The work approach adopted by the OCFL editorial group and the community contributions (use case submissions) made it so that libraries are available in Java, JavaScript and Python.

OCFL is a simple arrangement of files and collections of files in the digital storage structure. The architecture places the AIPs directly in the different layers of the system. Therefore, a deterministic mapping is necessary to precisely locate a specific file. The pairtree algorithm creates a file path by pairing successive characters two at a time in the file identifier. The use of a pairtree-like system maps a file into a folder hierarchy, and vice versa, and permits interoperability. Other OCFL requirements are the inventory.json and the conformance declaration file at the root of the hierarchy, which are both easy to carry out.

To achieve completeness, "*OCFL recommends storing metadata and the content it describes together so the OCFL object can be fully understood in the absence of original software*." (Hankinson *et al.* 2020, p. 1). This method for metadata handling avoids a supplementary operation metadata governance wise. The latter would be managed with content files. As a

result, since OCFL has no incidence on the nature of the file content, metadata format is also up to users' discretion.

The basic functions of persistent storage (that are "create", "read", "update" and "delete" or CRUD) as seen under the lens of computer programming (MDN [n.d.]), could be handled in OCFL:

- Create and Read: in a simple case where the system does not provide versioning;
- Create, Read and Update: the minimal functionalities in a system that supply versioning;
- Delete: would be an available option in combination with the above instances. It is worth mentioning that file elimination in a new version does not remove it from earlier versions.

3.4.2 Insights from preservation practitioners

We conducted three expert interviews to assess our conception of information architecture as well as obtain broader perceptions to be able to come up with the most comprehensive specifications. The more detailed they are, the easier it will be to ascertain their compatibility with DNA technology. Below are the ideas that came about during those conversations. They will be used to enrich our initial analysis.

Regular error checks represent a tremendous constraint on systems, even though they are highly recommended by OAIS. The same goes for the relevance of multiplying copies in the case of DNA medium. If there are many copies on a stable medium (physical redundancy), such as DNA, full checks are seldom necessary.

Considering the deficiency of the process in terms of reactivity, DNA storage would be better suited as a solution for disaster recovery. Even then, the recovery time must be taken into consideration: it must be adapted to how the institution functions and enable current operations to resume as fast as possible. In the case of a cantonal or state archive, a few, approximately two to three, months may be acceptable, in particular, if the order of restoration can be monitored: most important files to be restored first.

OAIS makes provision for AIP updates; it just does not supply the how. In practice this means managing the versions of data within the AIP, which implies:

- The need to prevent incidental data deletion;
- The importance of keeping the old versions;
- The obligation to unequivocally identify the different versions from each other (for reference or when directly quoted).

At the CERN, AIPs are generally composed of one file with its metadata. Files structure and versions are managed by means of AIC (Archival Information Collection), which are collections of AIPs. AIPs are immutable. This BagIt based system uses timestamp+md5 as identification and has the advantage of lightening the size of an AIP.

Using file checksums as identifiers would be advantageous: it would guarantee their uniqueness and avoid confusion problems.

It will be essential to include long-term and complete documentation independent of DNA technology, thus kept outside the storage system. This should include:

- Relevant documentation for information about the entire preservation system;
- The structure of the storage hierarchy (algorithm,...);
- The AIP standard;
- The metadata schema;
- The choice of DNA media implementation (even if it is a technology with a very low risk of obsolescence).

In terms of technical and volumetric considerations, experts recommend:

- Using the Unicode standard (e.g. UTF-8) would be ideal to be able to render all the characters and to allow a versatility and an exhaustiveness of the system.
- Planning for a large storage capacity: in fact, the volume required for a cantonal archive has an average size one Petabyte without the audiovisual data. It would be necessary to foreplan for the near future a system that can accommodate up to 10 Petabytes (without audiovisual).
- Guaranteeing authenticity, in particular the fixity of metadata, which is still little practiced.
- Revising metadata so that they can document the work of archivists in the way of a journal log.
- Using checksums as unique identifier for AIPs could be applied to DNA Example: SHA256(aip_id/version)
- Devising a formula that could provide an estimate of stored DNA shelf life based on DNA half-life.

  “Half-life” is a concept used in physics and biology that gives the time required for half of the amount of a substance to decay or be eliminated.

  Information carried by DNA will be accessible as long as enough DNA persists. This readability threshold is the function of DNA half-life since on reaching half- only half of DNA copies will remain. Half-life can be indicative of the time interval for which a complete check and correction is unnecessary. DNA is to be resynthesized once the limit of the number of copies for information accessibility is reached.

As for the format choice, experts recognized that OCFL is the most recent method to lay files on discs. However, it would certainly be interesting to consider other protocols, packages formats and distributed file systems.

The OCFL standard is new and still very untested. Chances are that the community will never adopt the standard the way they embraced BagIt which is widely used and whose libraries are more mature. What is more, the versioning proposed by OCFL poses a problem at the archival level. One approach would be to consider that AIPs are immutable to avoid any risk of confusion during citations. OLOS, a preservation solution that focuses on the reproducibility of research, chose that policy. Pragmatism dictates that there should only be one "good version". The relationships between versions can be described in the metadata by means of references between AIPs. It would be advisable not to be limited to what is offered by OCFL but consider the benefits of BagIt.

3.4.3 AIP specifications toward DNA storage

After meeting with each key informant person, we adjusted our specifications for storage and AIP structure. The following tuned results were then presented to the DNA applications specialist.

Due to DNA very low latency, we had to settle that, for the time being, DNA was not as efficient as current flash media. Therefore, we proposed a few uses cases for DNA persistence:

- As a backup copy for valuable records;
- As a data repair tool after error check;
- As a disaster recovery protocol supporting system;
- As a storage for hardly accessed but voluminous data.

Then, DNA would be a medium among others, as an addition to more time-efficient ones in a distributed digital preservation system for instance. It would be an offline medium stored in a secure location, safe from hacking, human errors or technical failure.

OCFL file arrangement with its application-independent approach and pairtree mapping would be used as a framework for digital information storage architecture. AIP would be managed as OCFL objects with the built-in checksum system that ensures fixity during storage and interactions with other media. Versioning would be an available option and CRUD operations will be conducted according to OCFL specifications.

As for error detection and correction, it would be more economical in time and financially to determine the period AIPs can go without being checked. Since DNA will be present in a certain number of copies, the system can tolerate a relatively long period between controls. Once the threshold for the minimum readable number of copies is reached, data is replicated into a new batch of DNA.

4. Digital archiving in deoxyribonucleic acid

4.1 Background

4.1.1 Deoxyribonucleic Acid

By 1954, scientists have discovered the main characteristics of deoxyribonucleic acid (DNA): its chemical components and structure, and its central role in the life machinery. It ensured the storing and transfer of information from one generation to another from the dawn of life. As such, it has developed outstanding features, making sure genetic material is stored very efficiently and transferred successfully.

Biomimicry has always been one of the most beneficial of human beings modus operandi. DNA ingenious way of naturally storing information makes it a perfect candidate for storing information.

"DNA sequences can contain more information than their binary counterparts because DNA with four bases has 4X representations possible for a character string of length X, while the binary system can contain only 2X times that information. In addition, data are stored in a volumetric fashion in a DNA molecule." (Panda *et al*. 2018, p. 3)

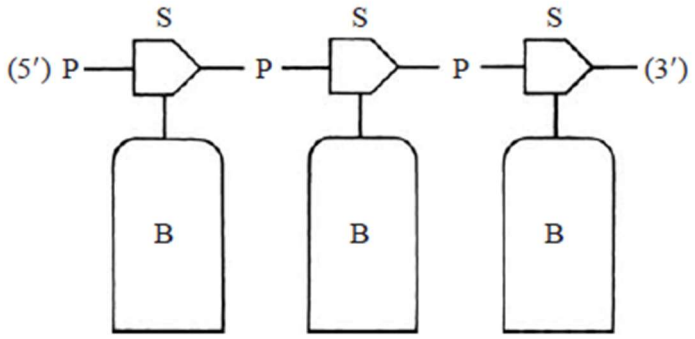
What is more, the amount of information DNA can contain in a very limited volume is remarkable.

"If we could successfully employ DNA to store data, the entire current global information (3.52 × 10^{22} bits) could be packed in a 0.00352 m^3 box and ~ 1 kg of DNA would be sufficient to address the world' storage requirement in 2040 (3 × 10^{24} bits)." (Panda *et al.* 2018, p. 3)

DNA is a slightly acidic chemical compound found in cell nuclei of living organisms. It is a nucleic acid. It is a polymer made of two twisted chains of nucleotides. Nucleotides are the monomers of each chain, DNA building blocks, and are composed of a base attached to a backbone made of sugar, the deoxyribose, and phosphate. There are four different bases:

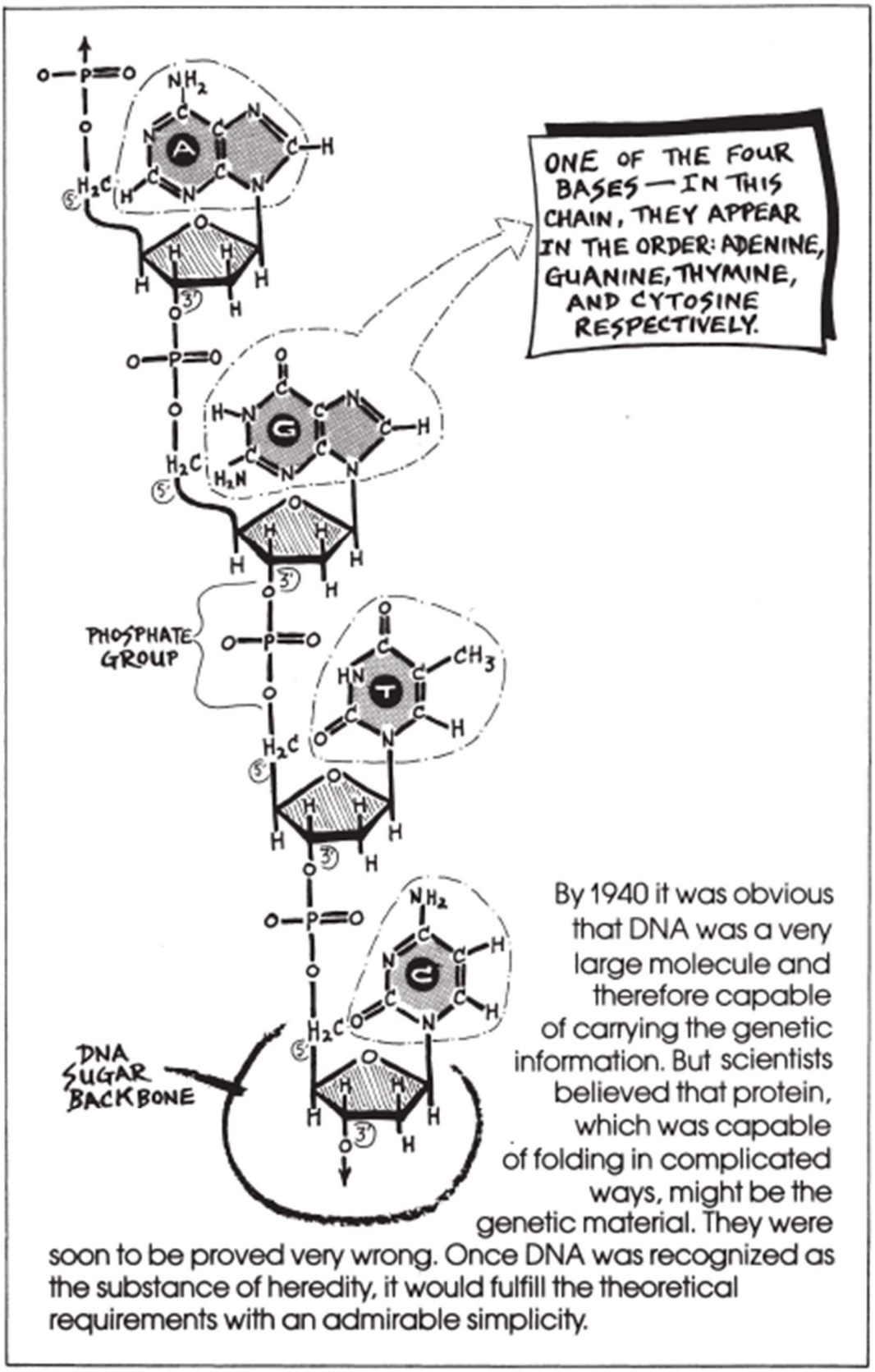
- Adenine (A)
- Cytosine (C)
- Guanine (G)
- Thymine (T)

Figure 13 : DNA thread arrangement showing the sugar (S), phosphate (P) and attached base (B)



(Calladine 2004)

Figure 14 : The nucleotides or bases, DNA building blocks (A, T, G, and C)



(Rosenfield, Ziff, Van Loon 2011)

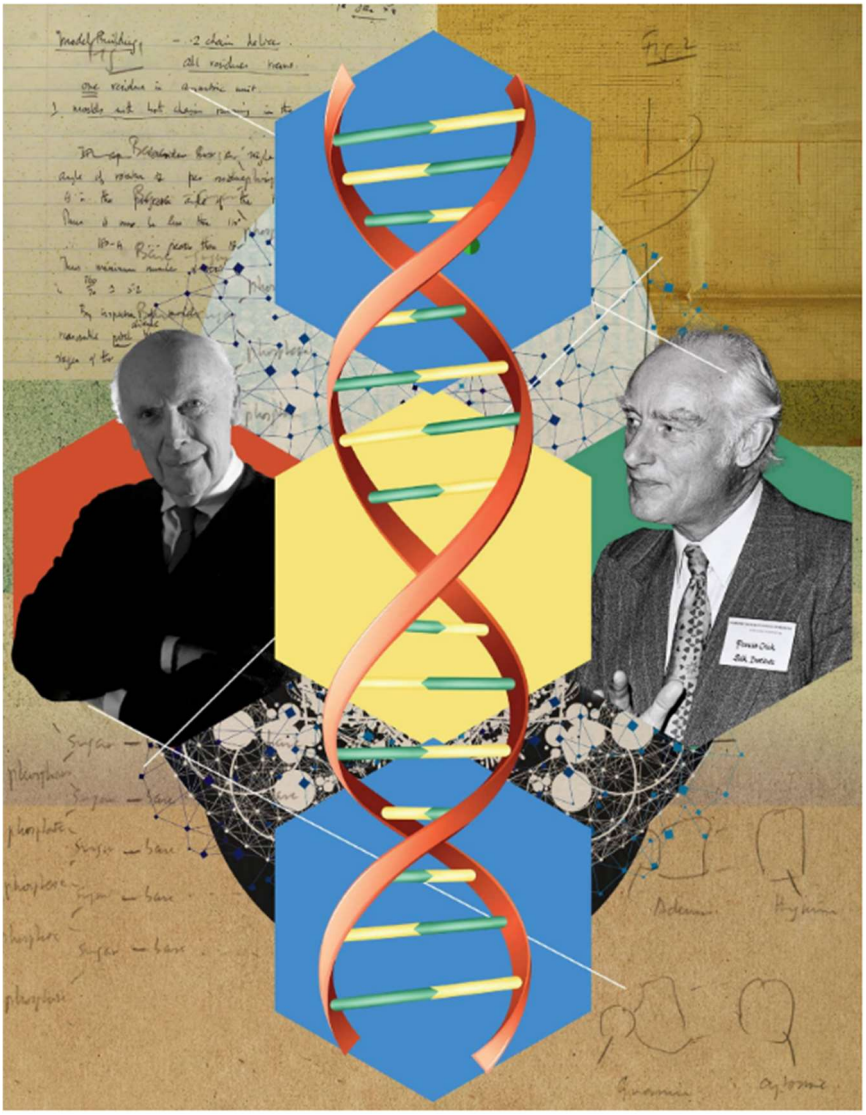
The way the bases are ordered in the DNA thread is responsible for encoding genetic information.

Hydrogen bonds between complementary bases maintain DNA two associated chains, known as the strands, together. Those complementary bases form a base pair (bp):

- Adenine pairs with Thymine
- Guanine pairs with Cytosine.

These pairings are essential to the three-dimensional double helix structure of DNA.

Figure 15 : Watson and Crick discovered DNA double helix structure based on Rosalind Franklin's work



(Weitzman, Weitzman 2020)

4.1.2 Storing digital information in DNA – Literature review

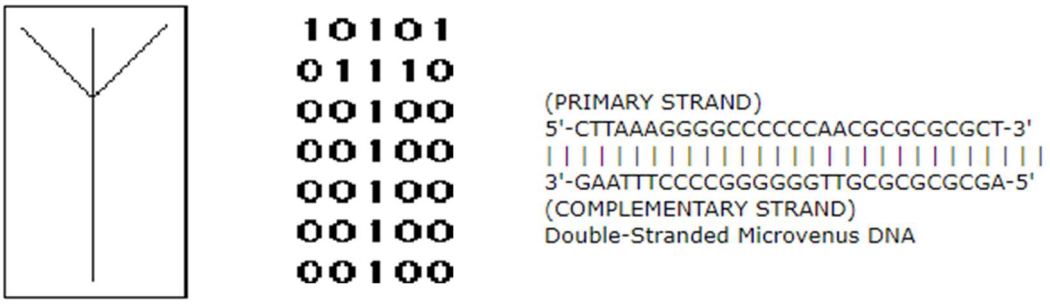
The use of DNA as a medium to store information did not materialize until early 2010 even though the idea has been around since the mid-sixties. The idea of storing information on DNA seemed to first occur to Russian physicist Mikhaïl S. Neiman while working on

microminiaturization in 1964 and then discussed with Norbert Wiener (Ceze, Nivala, Strauss 2019).

In the following section are a few milestones in DNA-based digital storage.

In 1988, with "Microvenus" Davis encoded an icon representing an ancient Germanic rune, “originally used to represent *life* and the female *earth*” (Davis 1996, p. 70), into a short strand of synthetic DNA (28bp). He then incorporated it into a bacterium. The transmission of the Arecibo message to space in 1974 inspired this work called a *living art* or *biological objet d’art*.

Figure 16 : Early work in DNA information storage, the Microvenus (1988)



(Davis 1996)

Clelland *et al.* demonstrated an encryption process using microdots in 1999. The experience consisted in encoding a secret message into a DNA sequence (69 nucleotides long with two primers – forward and reverse – of 20 bases each), synthesize it, then conceal synthesized DNA in a mix of fragmented human genomic DNA. A sample was extracted and reduced into a microdot resulting in a “double stenographic technique” (Clelland, Risca, Bancroft 1999, p. 533).

The previous work is the only one up until 2012 that does not involve any *in vivo* steps (Ceze, Nivala, Strauss 2019). In the present work, we are going to stick with entirely *in vitro* procedures. After 2012, DNA-based data archiving research seemed to gain more interest.

In 2012, Church and colleagues successfully encoded a 659 kB book into several short DNA strands (oligonucleotides library). Reading was then done through amplification (by PCR, explained on the 4.2.5 section) followed by library sequencing (Church, Gao, Kosuri 2012).

In 2013 Goldman *et al.* succeeded in storing 739 kB of data (text – The Watson and Crick 1954 paper on DNA three dimensional structure, images, sounds) using a compression algorithm (Goldman *et al.* 2013) resulting in the first compressed data storage in DNA.

In 2015, Grass and his team used the Reed-Solomon code to detect and repair errors (Grass *et al.* 2015). Even though data size was not significant, the use of this particular algorithm allows the compensation for the DNA-related errors happening during storage resulting in an error free information retrieval thanks to the application of the error-correction algorithm they were the very first to use. In the Grass experience, the DNA, protected by a glass bead, was heated at 70°C for one week. DNA aging simulation is performed by heating the sample. “*This*

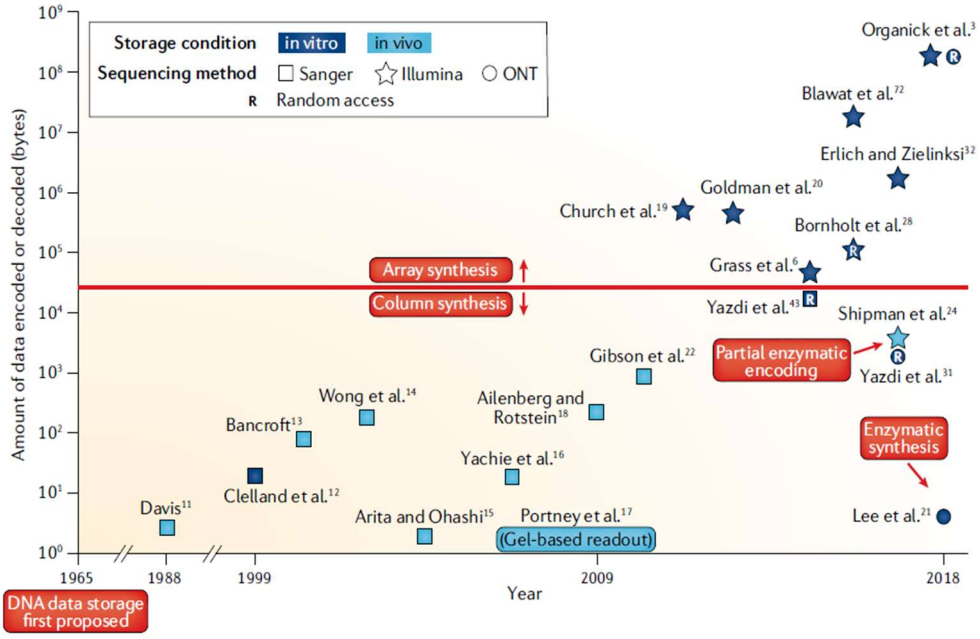
is thermally equivalent to storing information on DNA in central Europe for 2000 years." (Grass *et al*. 2015, p. 2552).

In 2017 using the Fountain code, Erlich and Zielinski report a coding potential of 1.98 bits.nt^{-1}, which nears the theoretical prospect of 2 bits.nt^{-1},and yielding an effective information density of 1.57bits.nt^{-1}. Moreover, the three steps Fountain code appears to be very efficient to correct errors since perfect data was retrieved (Erlich, Zielinski 2017).

In 2018, Organick *et al.* managed to retrieve individually each of 35 files stored in more than 13 million DNA segments (Organick *et al.* 2018).

In 2019, Lee *et al.* use enzymatic synthesis, which significantly reduces costs and allows for higher volume production of molecules (~500-1000 bases per synthesis) (Lee *et al.* 2019).

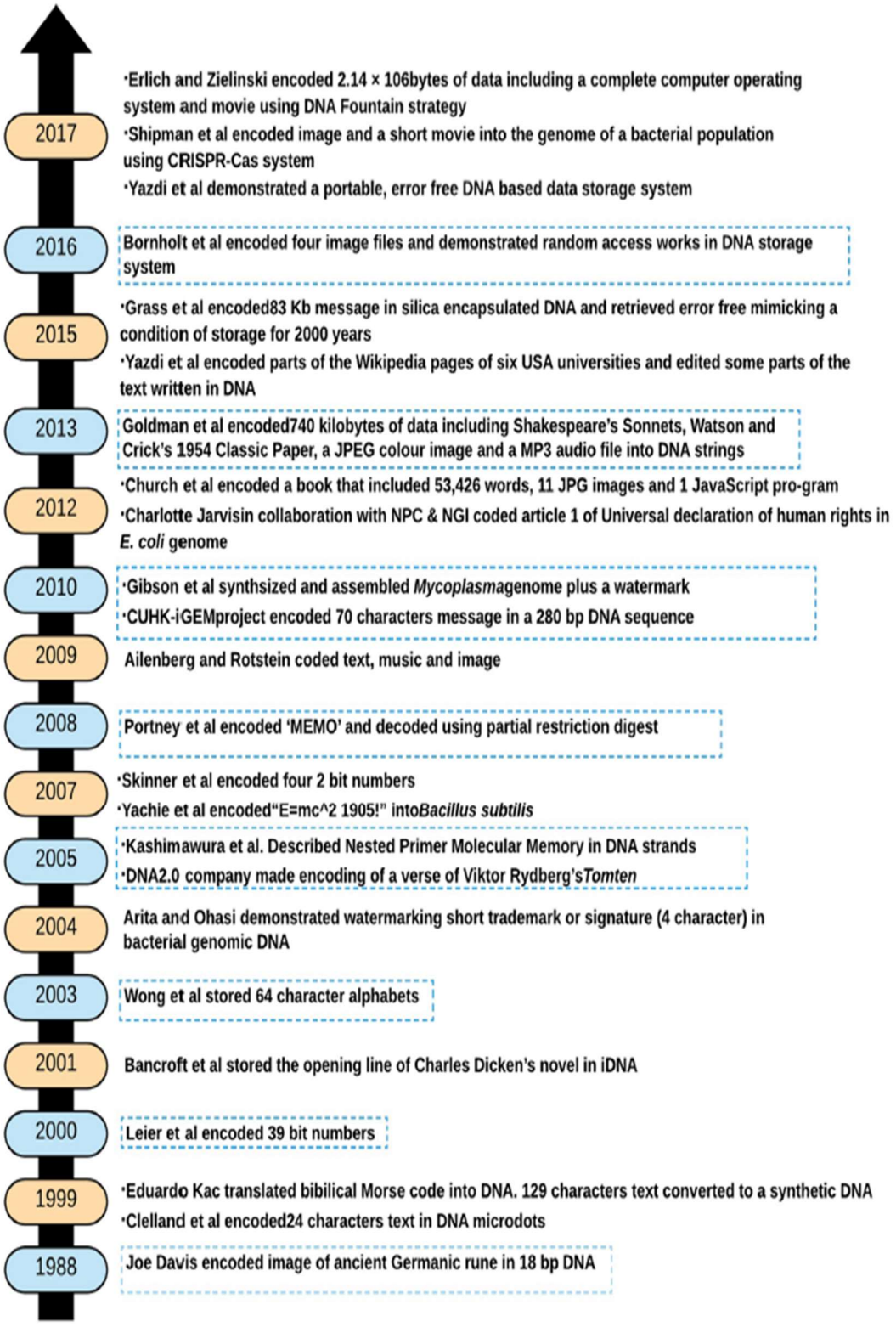
Figure 17 : Technical informative timeline of DNA-based storage experiments



(Ceze, Nivala, Strauss 2019)

Even though we will not take interest in *in vivo* experiments, they were quite popular before 2012 and deserve to be mentioned. The following section gives an overview of the main works done in the DNA-stored data fields between 1988 and 2017.

Figure 18 : More comprehensive timeline tracing key dates in DNA information storage

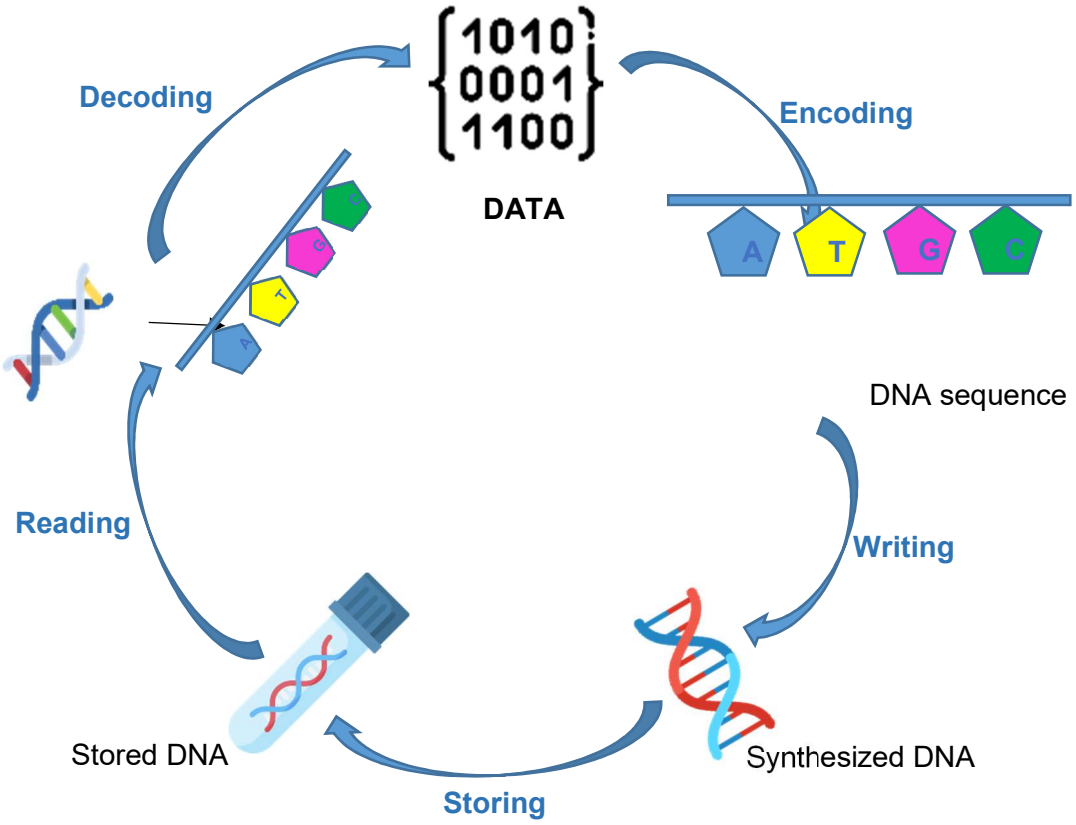


(Panda *et al*. 2018)

4.2 From digital information to storage on DNA

Figure 19[2] shows the major steps required for DNA digital information storing:

Figure 19 : The process of embedding and retrieving data from DNA



4.2.1 Encoding

This part of the process consists in converting the digital data into a DNA sequence. The very first codes used just did that job. Later, error detecting and correction were considered when choosing a code since DNA synthesis, storage and sequencing usually generate errors. The used code defines the quality of retrieved information. What is more, error-correcting codes add more sequences to the initial content. This induces higher costs, which always stand as a determinant factor.

The decoding is just taking the encoding step in reverse. This implies that the encoding part needs to be documented for the decoding to happen and for information to be correctly and fully extracted.

[2] The icons were designed by:
FREEPIK, no date. Icons. *Flaticon* [online]. [Accessed 13 July 2021]. Retrieved from: https://www.flaticon.com/search?word=DNA&type=icon

In the next section, we will see a few coding techniques used to turn digital data into DNA sequences.

4.2.1.1 Simple code

Used by Church *et al*. (2012), it stored ~65MB of data in ~8.8 megabases of DNA. Each strand contained 159 nucleotides. This approach uses the binary code where 0 is replaced by an A or C and 1 is replaced by a T or G for an effective information density of 1bit.nt^{-1}.

4.2.1.2 Huffman code

Used by Goldman (2013), this technique, developed by David Huffman in the 1950s, uses an algorithm that transforms each byte into ternary digits to avoid repetitions of motives that generate reading or writing errors.

4.2.1.3 Improved Huffman code

Used by Bornholt *et al.* (2016), this principle improves the previous encoding (use of trits) by adding another layer of cipher to prevent successive apparitions of the same nucleotide.

4.2.1.4 Reed-Solomon code

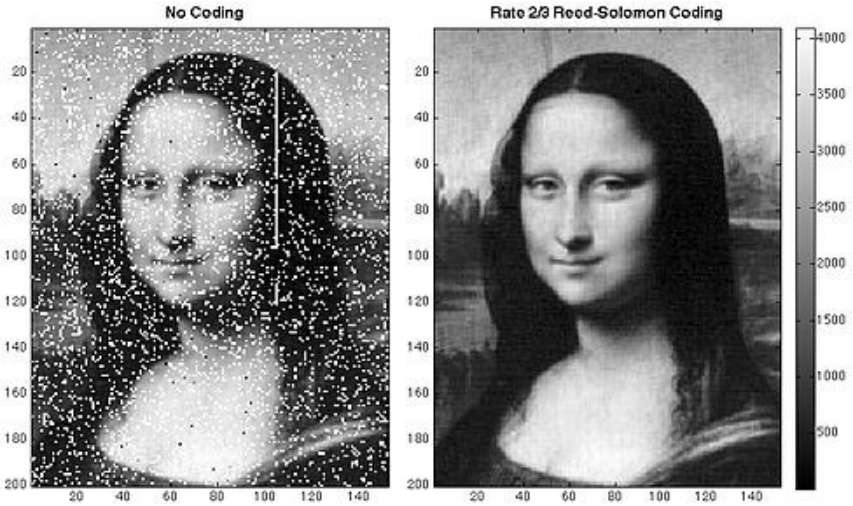
Grass and colleagues proposed in 2015 an encoding based on a combination of the Galois field and the Reed-Solomon code. This technique is particularly popular for error correction. It has been used in numerous commercial data storing devices (CDs, DVDs, hard disks, QR codes) and data transfer protocols (satellite communications, ADSL, xDSL). The Reed-Solomon code adds extra bits to digital data during storage or transfer to compensate for any error that occur (damages, interference,…).

Reed-Solomon codes, noted RS(n,k) can bring correction to t symbols according to the formula: $2t = n - k$

Where: k represents the size of the actual data

n represents the total size of signal (data associated with extra bits for correction)

Figure 20 : Error correction of Reed-Solomon code



(Goddard 2013)

In the DNA instance, it introduces a logical redundancy that allows information retrieval in spite of damages to fragments or loss of bits.

4.2.1.5 Forward error correction (FEC) code

A technique that improves the reliability of data by adding redundancies. The data is divided into blocks to which 3 nucleotides are assigned according to strict rules that prevent the occurrence of errors specific to DNA synthesis (insertion, deletion, substitution). The estimated encoding density is 1.6bits.nt^{-1} (Blawat *et al.* 2016).

4.2.1.6 Fountain code

Used by Erlich and Zielinski (2017), this code is reported to be robust and efficient. Encoded packets are generated from original data. Packets are made of a segment extracted from digital data (droplet) and the tag (seed) that identifies it. Packets are screened for repetitive nucleotides. Homopolymers containing sequences are discarded. Packets are produced until the length of original data (plus 5-10%) is reached. The Fountain code yields much lower redundancy and information density is the highest demonstrated at 1.98bit.nt^{-1}. However, reservations were expressed regarding its actual efficiency (Ping *et al.* 2020).

4.2.2 Writing – DNA Synthesis

The writing part corresponds to a DNA synthesis: the encoding produces a DNA sequence that needs to be synthesized in order to store it physically.

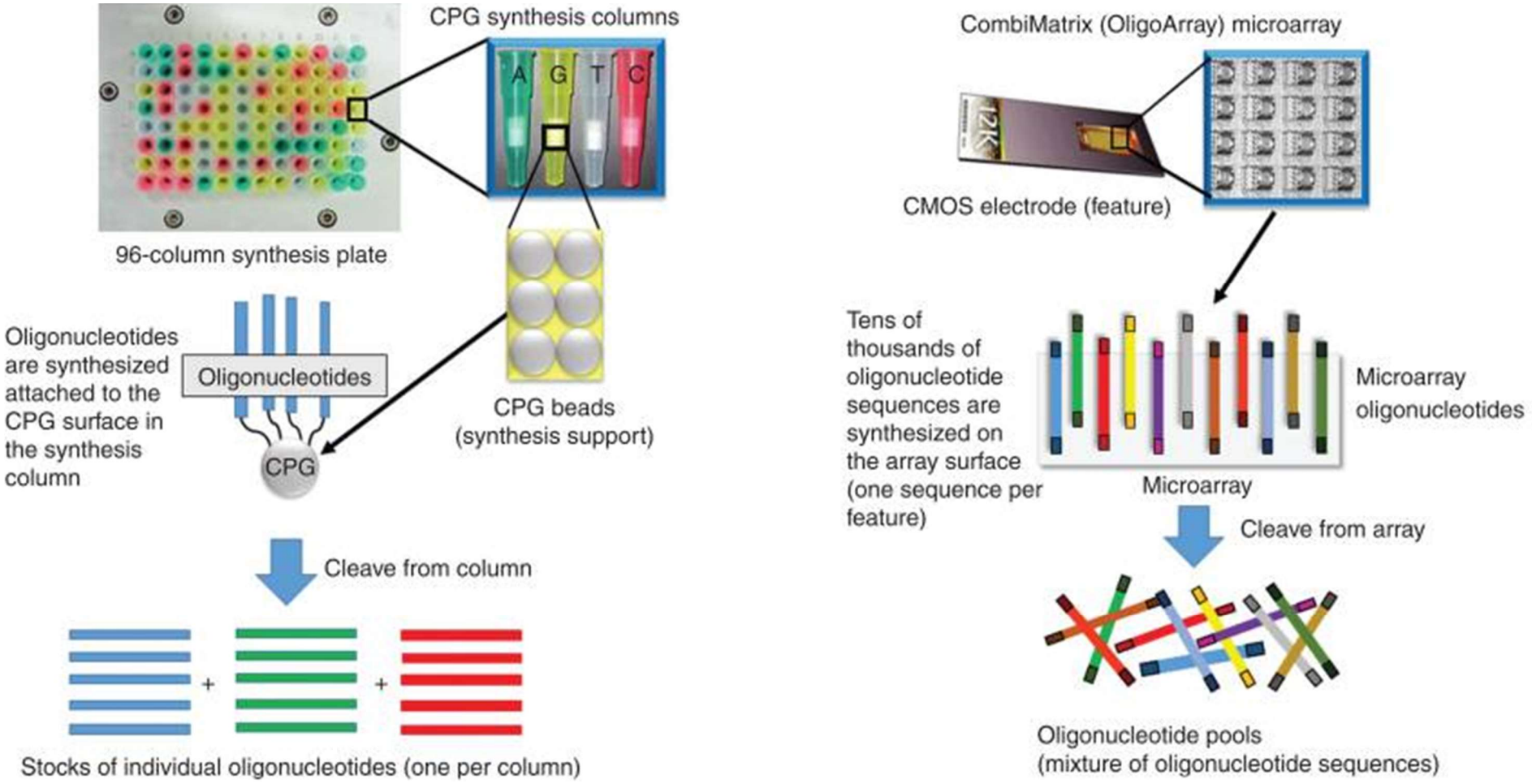
De novo DNA synthesis is indispensable to DNA-based data preservation even though it consists of a major bottleneck. The process itself is rather accurate. However, the slight error percentage grows exponentially with sequence length. Therefore, synthesized DNA length is technically limited to 150 to 200 bp. As a result, it is necessary to produce numerous small fragments, which implies costs. Most of DNA synthesis processes involve the use of phosphoramidite chemistry. Phosphoramidite is a chemical compound that can attach a nucleoside (a combination of a nucleotide and the sugar in the DNA backbone). As such, phosphoramidite nucleoside was used as the basis where polymerization starts in nucleic acid *de novo* synthesis.

- Column-based (low throughput) *de novo* DNA synthesis generates one DNA strand from a phosphoramidite nucleoside molecule.
- Microarray-based (high throughput) *de novo* DNA synthesis: using microarrays allows the simultaneous production of several different DNA strands on a static surface. In practice, manufacturers use equipment that can synthesize 1-9.10^5 unique strands in parallel.

It is currently still very challenging to obtain DNA strands of more than 200 nucleotides synthetically due to the accumulation of errors. It is easier to generate short strands but synthesis costs are certainly a bottleneck as already mentioned. What is more, error-correction codes introduce logical redundancy that summarizes the information allowing its recovery in the event of errors. These difficulties make research laboratories usually outsource DNA synthesis. The estimated costs are:

- $0.05 - $0.17 per nucleotide for column-based oligonucleotide synthesis (Hughes, Ellington 2017);
- $0.00001 to $0.0001 per nucleotide for microarray-based oligonucleotide synthesis (Hughes, Ellington 2017).

Figure 21 : Phosphoramidite chemistry – Column-based and Array-based DNA synthesis



(Hughes, Ellington 2017)

Lately, the use of enzymes is another way to generate longer strands of synthetic DNA, a method that brings the hope of reducing the synthesis time (40 seconds per cycle compared to 245 seconds per cycle for phosphoramidite chemistry-based synthesis) and costs *(*Lee *et al.* 2019).

4.2.3 Storing DNA

Currently, storing digital information is resource-intensive. DNA samples are usually frozen or chilled but that would consume energy. One of the main advantages of using DNA is the potentiality of room temperature storage for a very long period. Several protective (against water and air) methods are available to keep DNA intact for longer period:

- Encapsulated in glass beads (Grass *et al*. 2015; Organick *et al.* 2021);
- Stabilized with salts (Kohll *et al.* 2020);
- Mixed in sugars (Organick *et al.* 2021).

DNA longevity is usually measured by mean of its half-life. An entity's half-life gives the time required for half of it to sustain decay. Organick *et al.* compared those different methods and concluded that the most resilient (from DNA-aging simulation by heat) ways for DNA persistence are the ones with physical protection (glass beads).

4.2.4 Reading – DNA sequencing

The reading part means to take DNA from storage, reconstitute its sequence by DNA sequencing to finally extract the required information.

To determine the information DNA is carrying, it is necessary to identify the order of nucleotides in the thread. It is called DNA sequencing. For the purpose of the present work, only Next Generation Sequencing (NGS) based on the widely used Illumina technology will be described in detail.

The first step consists in an amplification of DNA by Polymerase Chain Reaction (PCR) performed to obtain a high concentrated DNA solution.

The second step is a DNA synthesis using tagged (by fluorescence) and protected nucleotides. This protection allows the attachment of just one nucleotide at a time. After the appending of each nucleotide, a signal is emitted and the wavelength of the dye allows the identification of the added nucleotide. Then the protection agent is removed so that the following nucleotide could be concatenated.

The third part is an analysis that compares all identified sequences to an original template.

As is done in DNA synthesizing, process parallelization can increase production.

Other DNA sequencing methods are:

- The Sanger dideoxy method that Frederick Sanger introduced in the 1970s. It has been automated since and can sequence more than "*[one] million bases per day*" (Berg *et al.* 2019, p. 546). It was the go to method until NGS equipment was marketed in the early 2000 and is considered low yield.

- Third generation sequencing, that uses nanopore technology, which bypasses the PCR step found in NGS, allows the reading of longer DNA strands (200kb) and an accelerated reading (10^{-7}-10^{-6}h.kb^{-1}) (Dong *et al.* 2020). Currently, its costs and error rate are still high.

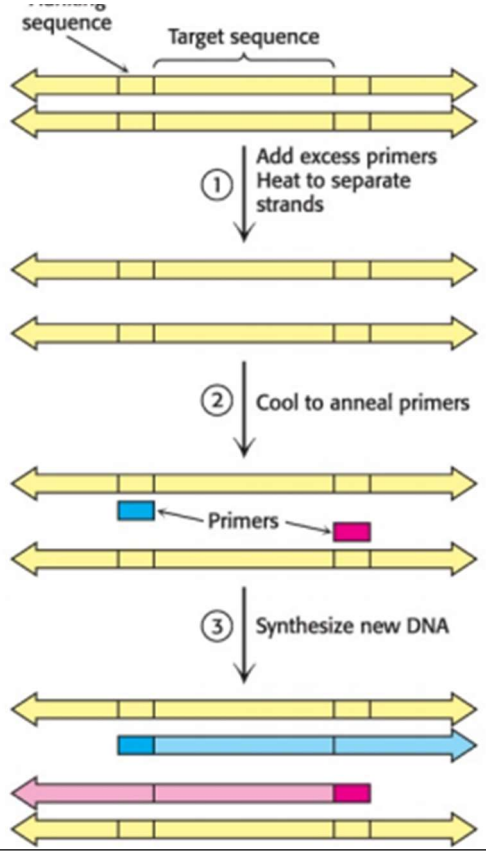
4.2.5 Access to information

Access function does not appear in figure 19. However, it is an unavoidable action before one can read the data intended for preservation. Polymerase Chain Reaction (PCR) enables access to information stored on the DNA medium. PCR is a DNA sequence selective amplification technique. Designed by Kary Mullis in 1984, it is a way of rapidly obtaining thousands of copies of a specific DNA sequence. It uses primers, single strand small DNA segments that allow the "localization" of a specific DNA strand containing a precise information by attaching to a matching region on the DNA sequence. The process involves three steps:

- Separation of the twisted chains into two single strands by heating the sample. It is the *denaturing* step.
- Attachment of primers at the end of each template DNA. Primers are short DNA sequences corresponding to the beginning of the target sequence. It is impossible to generate synthetic DNA without the aid of this starter. Its French denomination "amorce" is more illustrative of its function. It is the *annealing* step.
- Synthesis of the new DNA strand by heating the solution at 72°C, for the nucleotides to append to the growing new strand thanks to a chemical agent called DNA polymerase. It is the *extension* step.

These three steps make one PCR cycle. At the end of the first cycle, two new DNA strands, complementary to the template strands, are made. After 30 cycles, the DNA sequence is amplified to a billion fold, which can be achieved in less than one hour (Berg *et al.* 2019).

Figure 22 : First cycle in a Polymerase Chain Reaction (PCR)



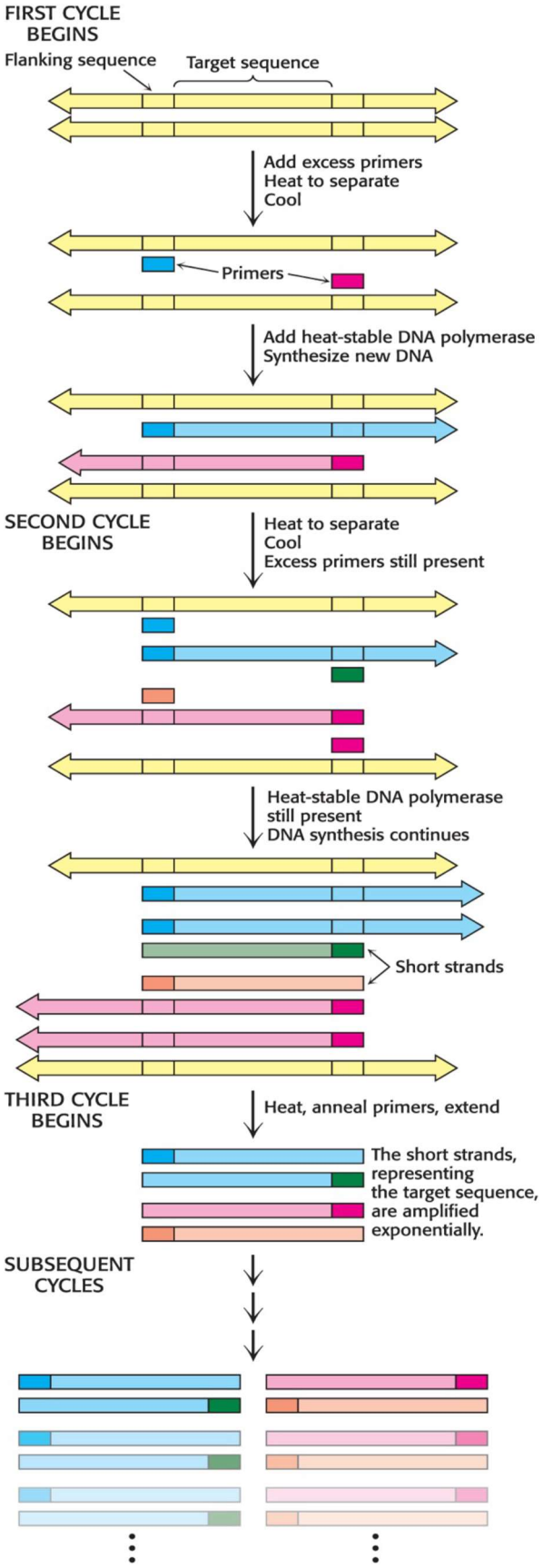
(Berg *et al.* 2019)

Primer design is a crucial step when doing PCR. Primers need to be complementary to the start and end parts of the sequence interest without matching the rest to prevent random hybridization. Four important points are to be considered when designing primers:

- The length: an average length for a primer is about 20 nucleotides (18-24 bases);
- The denaturing and annealing temperatures;
- The percentage of G and C relative to the length of the primer: G and C make stronger bonds that will facilitate annealing and extension;
- The secondary structures that are any other hybridization, besides DNA sequence-primer pairing, that can happen in the PCR mix and are to be avoided. The primer can bind to another primer or bind to internal structure present in the primer sequence.

Many primer design softwares exist on line. However, primers efficacy must be confirmed experimentally by generating a standard curve. A standard curve is obtained by regression after plotting the results of PCRs with different dilutions on DNA concentration and quantifying the amplification rate. This shows that primer design cannot be entirely automated.

Figure 23 : Description of the three first cycles in a Polymerase Chain Reaction



(Berg *et al.* 2019)

4.2.6 Perspectives in DNA biotechnology

Technological advances in synthesis and sequencing are more than likely to reduce costs and latency.

De novo DNA synthesis using enzymes looks promising (Lee *et al.* 2019; Dong *et al.* 2020):

- Synthesizing longer strands (about 200 to 1000 bases per synthesis) seems possible;
- Synthesis costs look already more advantageous than phosphoramidite chemistry: enzymatic DNA synthesis uses fewer reagents and the enzyme terminal deoxynucleotidyl transferase (TdT) can be used several times;
- The enzymatic reaction is six times faster than the chemical reaction.

The error rate is still high (~30%) and the whole process still has room for improvement mainly in its miniaturization and automation.

Third generation sequencing technologies are currently being tried and tested and seem to be on track to replace NGS, especially nanopore sequencing. Powered by Oxford Nanopore Technologies (ONT), it exploits the possibility of measuring current variation generated by the processing of a single nucleic acid strand through a pore. Current level detected within the pore differs from one nucleotide to another, which identifies them in the sequence. This technique allows the sequencing of longer strands in less amount of time. However, error rates are still high though costs appear promising.

4.3 DNA as a medium for digital data storage

We met (virtually) with a researcher who specializes in DNA applications and asked about the applicability of DNA technology to the specifications we established previously.

Other technical issues we had to inquire about were:

- The possibility of using Unicode standard for file names and the limits of file paths;
- DNA potentiality and the high storage capacity (in the Exabyte range) required by institutions such as state archives or research data repositories;
- How DNA latency could be addressed: what are the highest speeds in writing and reading in DNA? What about the costs?
- Could DNA shelf-life be estimated mathematically?

4.3.1 Storage layout feasibility

OCFL file layout does not have any incidence upon AIP storage onto DNA technology. The only hurdle is the necessity of fragmenting the encoded DNA sequence into strands of limited length. Directly encoding the AIP path and the AIP (unique) identifier in the primer is doable. However, one must take into consideration the limited primer length (about 20 nucleotides) when using hashing algorithms to generate an AIP identifier or AIP path in the file system hierarchy. They would be encoded as binary files thus encompassing the use of Unicode for file names. Using the identifier as a basis to design the primers would allow directed file pick up regardless of file layout. A difference in the “depth” of primers would allow determinant retrieval whether only an AIP is needed or the whole collection wanted. However, a primers

library would have to be kept separately as a written record. If primers list happens to be lost, the whole library could be obtained back after sequencing the data.

4.3.2 Data capacity

The theoretical limit for DNA storage is 17 Exabytes per gram (at 10 copies physical redundancy) but there are various limitations. For instance, synthesis is certainly a bottleneck (mainly at costs level) but there are other parameters such as redundancy (logical and physical) to consider.

4.3.3 Versioning

With different series of designed primers, versioning could be achieved since they would guarantee that different versions could be discriminated against each other. Encapsulated DNA could be added to a tube already containing previous versions. However, file updates such as is recommended in the OCFL standard is inefficient and should be avoided.

4.3.4 Storage

Silicate beads encapsulation is the preferred route for the Grass research group since the technology is available since it was developed by the team (Paunescu et al. 2013). It is a relatively strong physical protection, as bone would be to DNA in fossils, and can guarantee its longevity. Hydrofluoric acid in small concentration is needed to extract DNA from its glass armor ensuring water resistance.

4.3.5 Latency

Outsourcing DNA synthesis is more practical. An estimated speed of 1 to 5 minutes per base is a possible expectation however subcontracting means having little impact on the workflow. Even though a unique strand can be synthesized in each well of the micro-array card, in the end, the number of strands really depends on the file size.

The most optimistic reading speed is estimated at 300 kilobytes per second without any physical redundancy (a strongly discouraged practice) performed with the most performant available benchtop sequencer.

4.3.6 Encoding and correction algorithms

Reed-Solomon is a tried and tested code in telecommunication industry. The algorithm is stable; the correction rate is significantly close to that of Fountain code. The same goes for the net information density of the two codes (1.78 bits per nucleotide versus 1.98 bits per nucleotide respectively).

Reed-Solomon entails more logical redundancy that makes it more expensive (adds more length that needs to be synthesized).

4.3.7 DNA-based digital preservation costs

Costs are still high. To avoid information handling by a third party, buying a sequencer would be a good thing. Even though it is a one-time deal, it is still costly. DNA synthesizing manufacturers have a price range of about 300$ for 200 kilobases.

DNA manipulations (PCR, sequencing, and encapsulation/decapsulation) require manual handling, meaning the archival institution will have to hire at least a laboratory technician.

5. Scenarios

In the following section, three scenarios of DNA-based digital information preservation that can happen in an archival system are given. They were proposed in order to better understand the execution of the operations. Beforehand, we give a description of AIP designation and identification. This step is important to design the primers later. The first example shows the detailed process of encoding and writing. The second is about making an additional copy of the whole collection in distributes preservation system policy or to refresh a medium. The third presents a case of a specific AIP request and how retrieval is achieved.

5.1 AIP description

All AIP have their own unique identifier (PID) according to the following schema:

- Possible schema: <namespace> : <id> / <version>
- Example: “acv : 1 / v1”

It is important to mention that the proposed schema is given as an example. Systems have their own fine naming structure.

For systems that do not allow updates, version number will remain “v1”. AIPs are compressed as ZIP files (as ZIP is well documented open standard suitable for long term preservation, generally reduces the size of data, and generally reduces homogeneity which is an advantage when writing DNA).

5.2 Use case 1: Encoding and writing

The first example is set for a case when an Archive wants to replicate AIPs and their associated metadata as additional copies stored in DNA.The data (content and metadata) first undergoes a randomization session (e.g. using the XOR operation) to avoid repetitive sequences that can impair DNA writing and/or reading. This will also ensure the ratio of GC content will remain within the acceptable range of 40-60% necessary to DNA stability.

Since DNA synthesis of large strands (<200bp) is still highly inconvenient, digital data is fragmented to later obtain short strands of DNA. Fragmented data is submitted to a first Reed-Solomon pass (parameters n1 and k1) generating the *outer code redundancy* that will allow correction in case of loss of DNA fragments.

Indices are added to each sequence to determine their initial order since the fragments are kept in a DNA pool.

The second layer of correction code is applied: a second pass of the Reed-Solomon algorithm (parameters n2, k2) is applied. This will allow the correction of a single or a few nucleotide errors occurring within the fragment (suppression or substitution of a nucleotide).

Finally, the digital data can be converted to DNA sequences by storing two bits in a nucleotide as follow:

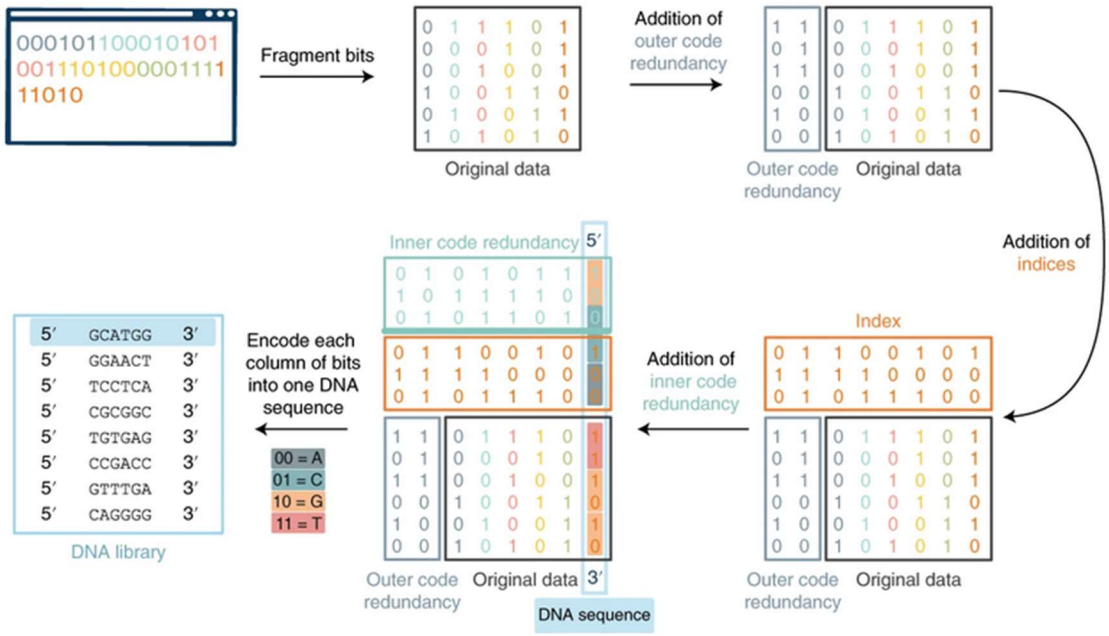
- A => 00
- T => 01
- G => 10
- C => 1

5.2.1 Procedure

We proceeded to an *in silico* simulation using the Python language. The whole process is described as following:

```
<dataC>    = reedSolomon(<data>, n1, k1)
<dataC2>  = [  reedSolomon(<indice1><dataC>[0 :N-1], n2, k2) ,
                 reedSolomon(<indice2><dataC>[N:2N-1] n2, k2)  ,
                 reedSolomon(<indice3><dataC>[2N:3N-1] n2, k2)  ,
… ]
```

Figure 24: From data to DNA – Detailed process



(Meiser *et al.* 2020)

For data retrieval, all the above steps are taken in reverse.

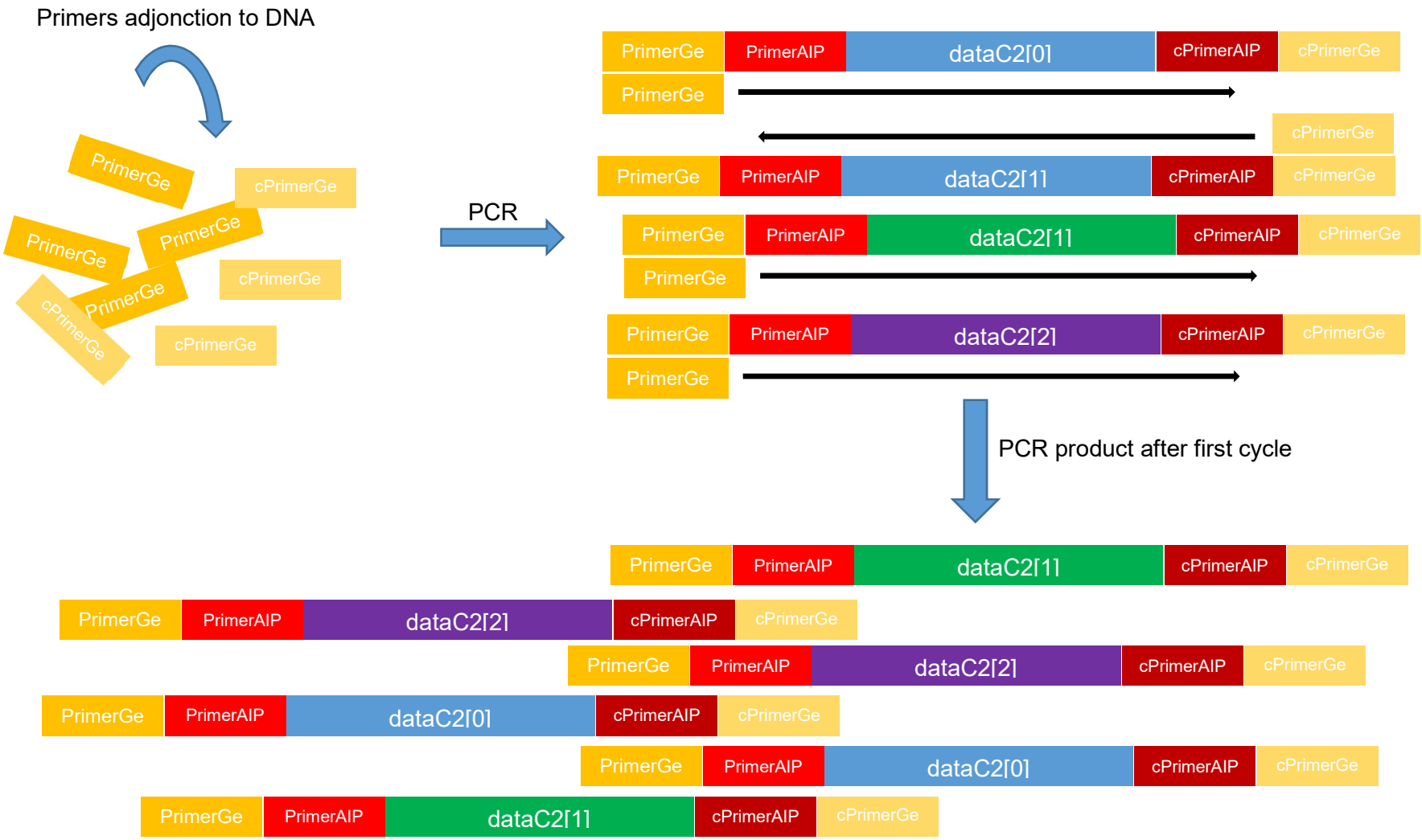
5.3 Use case 2: Making copies of the Archive collection

In order to make multiple copies of the entire archival collection, a PCR is conducted using primers that we will name **General primers**.

- General primers attach to a sequence region common to all DNA fragments in the DNA pool. This hybridization is the basis for a rapid replication of the Archive content creating a physical redundancy.
- They are designed specifically, using the namespace (common to the whole collection) as basis.
- Structure: PrimerGen = hash(<namespace>)[0:n-1]
  where n represents a defined number of first bytes
- Example: PrimerGen = sha256(“acv”)[0:11]

If the hash is expressed in hexadecimal, 12 “characters” correspond to 24 nucleotides.

Figure 25 : Replication of the whole collection



5.4 Use case 3: Directed AIP extraction

The third example concerns the need to access a particular AIP. Upon request for a specific AIP, PCR is used to locate and amplify the sequence of interest.

To discriminate individual AIPs within the collection, a second type of primers, that we will name **AIP primers,** are used. These are the AIP primers characteristics:

- Specific to an AIP, they are used to "locate", by hybridizing with a sequence specific to the DNA segments composing the requested AIP, and amplify it for retrieval.
- They are designed using the AIP unique identifier (PID) as basis;
- Structure: PrimerAIP = hash(<PID>) [0:n-1]
  Where n represents a defined number of first bytes
- Example: PrimerAIP = sha256("acv : 1 / v1")[0:11]
  Which corresponds to 24 nucleotides.

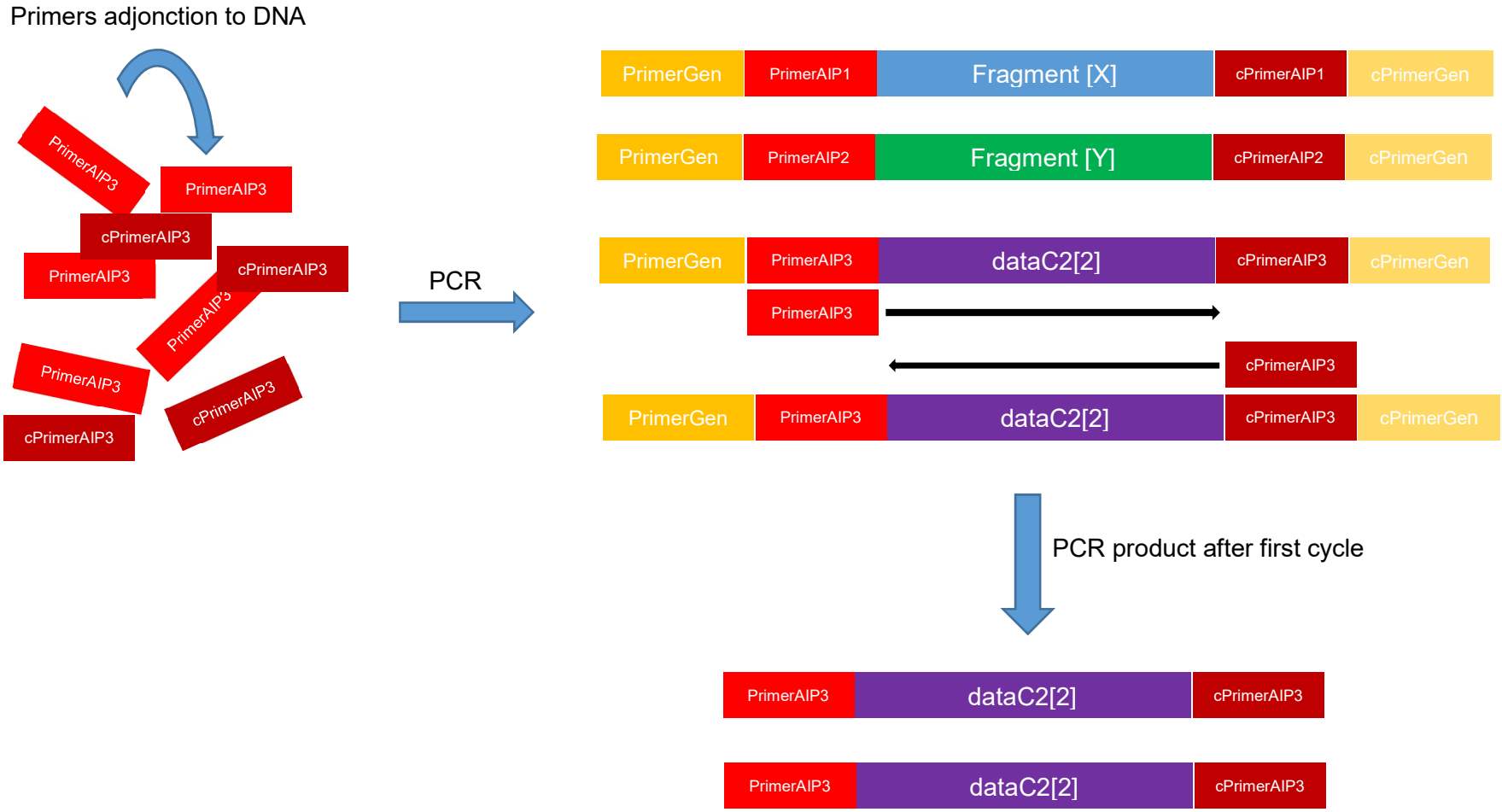
5.4.1 Procedure

A simulation would give:

FragmentsAIP = [<PrimerGen><PrimerAIP1><[Fragment>[X]>))<cPrimerAIP1><cPrimerGen>,
<PrimerGen><PrimerAIP2><[Fragment>[Y]>))<cPrimerAIP2><cPrimerGen>,
<PrimerGen><PrimerAIP3><dataC2>[2] >))<cPrimerAIP3><cPrimerGen>,
…]

Where <cPrimerGen> and <cPrimerAIP*> are the reverse primers for <PrimerGen> and <PrimerAIP*>. Primers are directly synthesized as small DNA sequences, so they do not undergo randomization or correction code. On the contrary, that is the case for <PID><data*> that were initially digital data subsequently transformed into DNA sequences.

Figure 26 : Targeting a specific AIP



6. Recommendations

The purpose of this study was to assess the feasibility of adopting DNA storage in an OAIS-compliant digital preserving system. Below are the few recommendations that could be drawn from this work.

6.1 OAIS compliance

According to OAIS specifications, OAIS-compliance in is achieved by:

- Acceptance of the information model as it is represented in chapter 2.2 of the ISO:14721 text;
- Meeting the mandatory requirement defined in 3.1. of same text.

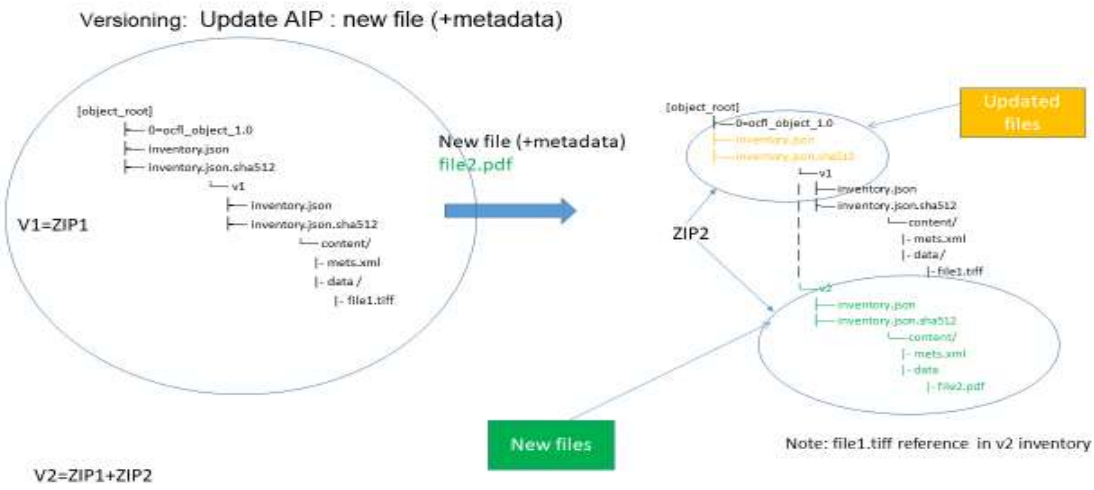
Using the OCFL standard as a basis for files organization is a simple and transparent way of administering and storing AIPs. The OCFL Object specification adequately conforms to designated AIP in OAIS standards. The parsability OCFL is striving for particularly address the need for preserved information to be understood without recourse to any specific software. The ability to rebuild the Archive is one of the founding principles of OCFL. As the evolution of Moab, which itself is an improvement of BagIt, OCFL aims to circumvent all the previous pitfalls. Digest for each inventory and a general checksum for the whole OCFL object guarantee information fixity. Combined with versioning this feature allows for content history to be traced.

The only drawback for OCFL implementation in DNA medium was the inefficiency of updating individual small files such as the general inventory and its checksum. However, it is possible to bypass this problem. The solution would be to associate the updated general inventory and its checksum in each newest version folder.

In the following example, the initial version “v1” would be zipped up in what will be named “ZIP1” in this instance, encoded then synthesized into DNA.

Versioning would create a second set of files that would be contained in another zip file, “ZIP2” in this instance. On top of its regular content (in green in figure 26), ZIP2 would include the conformance declaration, the AIP inventory and its digest (in yellow in figure 26). This would require the opening of the latest version to consult the AIP’s inventory.

Figure 27 : Handling OCFL requirements on DNA medium



6.2 Digital archiving on DNA

In this part of the work, we are going to examine each step of the DNA preservation procedure and determine the go-to course of action based on current practices on DNA preservation.

6.2.1 Encoding

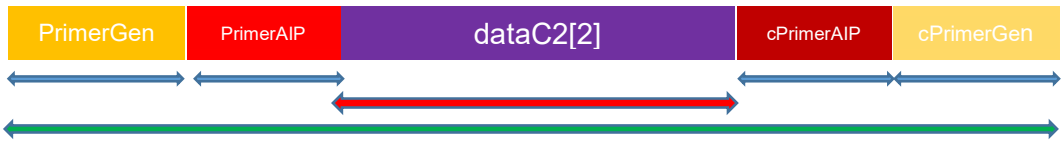
We have listed several encoding methods in section 4.2.1. of present work. All of them have been used in a study involving using DNA for digital information storage. Despite Fountain code asserted efficiency, relying on the Reed-Solomon code would be more recommendable. It is a robust and corroborated error-correction algorithm.

Primer design will require special processing since it cannot be fully automated for the moment, especially for synthesized DNA.

6.2.2 Writing

As stated, *de novo* DNA synthesis is complicated and more often than not, laboratories simply outsource DNA synthesis from three main manufacturers that generally practice the same ratings (costs and time) with very high synthesis precision. Resorting to suppliers, however, raises the concerns about having data handled by external entities and having no incidence on the production rate. In the end, costs remain the limiting factor since AIPs would be fragmented into numerous (from tens of thousands and up) short strands including physical redundancy.

Figure 28 : Synthesized DNA length



In the above picture, the total length of DNA to be synthesized is represented by the green arrow. It is limited to about 150bp. The blue arrows give the length of primers, about 20 nucleotides. We have two primers “PrimerGen” and “PrimerAIP” and their respective reverse primers “cPrimerGen” and “cPrimerAIP”. Which makes about 80 nucleotides to be subtracted

from the initial 150 nucleotides. The red arrow represents a length of about 70 nucleotides dedicated to the actual data. And as mentioned before, the AIP fragment is subjected to the error-correction code. In the end, a very small portion of synthesized DNA length actually carries the information. This is where the need to generate many fragments stems from.

6.2.3 Storing

Mimicking DNA preservation in fossilized bones, the use of glass nanoparticles to encapsulate DNA appears the best way to keep them intact the longest at room temperature. However, the encapsulating process requires a few days (6 days) to happen including a few hours (about 4 hours) of manipulation (Paunescu *et al.* 2013; Meiser *et al.* 2020).

The reverse action, liberating information from nanoparticles, takes about 10 to 20 minutes by dissolving the silicate in a diluted hydrofluoric acid solution. This step must be done under a chemical hood, which the ACV are already equipped with. Physical redundancy allows the extraction of amount of DNA for reading purpose and the remaining material can be encapsulated once more.

6.2.4 Reading

PCR and NGS are the usual pathways to access and retrieve information stored on DNA. Thus, storing digital information on DNA entails a few actual and unavoidable human interventions. In the case of an archival institution such as the ACV, one of two different approaches can be adopted:

- The first option would be to set up an in-house small-scale laboratory which means purchasing an equipment (onetime acquisition) and hiring a laboratory technician (annual expenses depending on their condition of employment);
- The second method would be to arrange for a collaboration with existing laboratories to handle those manipulations.

In both cases, available funds will be determinant: synthesis costs will still constitute a serious bottleneck. The criteria for selection would be the Archives' needs and throughputs in terms of amount and flow of information.

In case the second route is chosen, the nature of the piece of information submitted to an external handling requires prior consideration: a threshold for sensitive data should be established. A workflow ensuring the integrity of data transfer with a guarantee of AIP safe reception and return must be thought up established and documented.

An in-house laboratory can ensure a control over every step of the process and free from the necessity to hand over data to a sub-contractor. To absorb the costs of setting up the laboratory and the charges for the hire, a steady production would be required. More thorough comparative studies between in-house handling and outsourcing, especially for the access and DNA-reading procedures, can be conducted to determine which method is most advantageous according to institution needs.

6.3 Documentation

To meet the completeness demanded by OCFL, it is mandatory to document in detail the entire process. The following list enumerates the minimum required items. Exhaustively documenting the work is essential for system to be auto sufficient and reliable for reconstitution.

The documents should be kept in paper form along with storage:

- General project procedure documentation
- AIP unique identifiers structure
- Primers design procedure (software, melting temperature, standard curve)
- Procedure for coding AIPs into strands and reciprocal procedure
- Code, libraries:
  - Should use a go-to language, if possible pseudocode (e.g. python)
  - Hash (sha256)
  - Library: Reed-Solomon, zip,
- Used calibrating parameters
  - Bit to DNA dictionary
  - Random data for the mask
  - Parameter of the interleaved ReedSolomon algorithms (2 parameters sets)
  - Data length for each strand
  - Primer length
  - Length of space to store identifiers

Laboratory protocols:

- Information accessing process
- DNA reading procedure

6.4 Proof of concept

The need to set up workflows and improve processes automation is obviously one of the main priorities of the field. Since this kind of work has never been done in any archival organism before, it is necessary to conduct a pilot in order to experiment in the working frame of an archiving system as a proof of concept. A pilot would aim at establishing a viable protocol from encoding to reading within an archival system.

A schema for primer design that considers the systematic handling of AIP unique identifier (used algorithm such as hashing function and the way it is applied) is necessary. The idea of producing different depths of primers (general and AIP-specific) demands a prior conception since primer length is limited but then the two different primers must be integrated into the DNA sequence with their respective complementary strands. It influences the final length of DNA to be synthesized since the effective DNA encoding for data of interest would be shortened.

It is also imperative to devise an efficient way to store physical DNA containers in the archival system and define how to easily locate the required file, which is stored in a particular vial on a specific shelf. Archival institutions already have a filing system for physical items storage. This puts Thibodeau's reflections about digital object embedded into physical objects in perspective (Thibodeau [n.d.]). It will be necessary to ingrate this new addition into the existing storing pattern.

It could be interesting to devise a mathematical formula that can predict the half-life as a function of the redundancies (logical and physical) and the storage temperature. But in a more

practical way, the pilot could integrate the testing of DNA half-life by running tests periodically (every one, two, five, ten and twenty years) to check information readability under real conditions. At the same time, a limit access rate (number of times the AIP can be requested) could then be determined from a fixed amount of synthesized DNA.

7. Conclusion

This work was done under the focused lens of Archival Storage Function entity of OAIS since its main goal was to identify the specifications needed to implement a novel storage media, DNA, in an archival system. OCFL lends its structures as a robust framework able to sustain a DNA based preservation storage. Wilson describes digital preservation as:

"*Bit-level preservation and appropriate object metadata are both necessary. Without both, a digital preservation system does not exist*." (Wilson 2017, p. 130).

The OCFL layout fulfills this role within the OAIS specifications and its simplified and application-independent architecture allows the application of DNA-based storage.

So, DNA archiving is realizable and a pilot project can be conducted with the findings from the present work. However, DNA-based digital preservation scalability is still on the low due to financial considerations, mainly DNA synthesis and sequencing are costly. Compared to currently used storage media, latency is still considered mediocre.

Certainly, the system will need improvement in order to be competitive. For DNA to become a staple in data storage, automation of the entire procedure must be reconsidered since many steps along the process still require hands-on manipulations.

Moreover, DNA reading, writing storing and even information access methods are currently being investigated and so far, the results are encouraging.

In October 2020, Antkowiak *et al.* proposed a low-cost synthesis method that yields a less precise DNA. This impediment can, however, be counterbalanced by the use of an optimized error-correction code. In addition, using DNA as a storage medium allows more latitude than biology would because the same precision is not required so the margin of error can be greater. We could therefore favor density over precision, a possibility that could save time and money. Along with enzymatic synthesis, this promising technique is undergoing further developments.

DNA storing methods need to be investigated more thoroughly under real conditions (storing method, temperature, physical redundancy, logical redundancy). Glass encapsulation is robust and stable. The density information it can contain is high, "*3.4 weight per cent of DNA to encapsulant*" (Organick *et al*. 2021), however, the process is time consuming and information retrieval can last up to 20 minutes. Mixing DNA with a sugar such as Trehalose is a less complicated approach but information density droops to 0.13 weight per cent and since the aging process was a simulation by heating, one cannot ascertain decay in real conditions.

DNA half-life at 15°C is estimated at 1000 years and it decreases to 100 years at 20°C. DNA aging experiments have been carried out by heating the sample however no one has ever tested DNA in real conditions for 1000 years. So, we do not know whether or not these predictions are accurate or if DNA decay behavior changes in real conditions. It would be necessary to try to preserve DNA in archives to get a real take on how time resistant DNA really is.

For information access, a random pick up technique that bypass PCR by labelling the silica containers that protect DNA have been tried (Banal et al. 2021). It offers the advantages of avoiding multiple heating/cooling demanded by PCR as well as a higher latency.

For DNA reading, nanopore technology sequencing is up-and-coming and has already been miniaturized. It has longer reads and lower costs and better speed. The error rates are still high but based on the trends of the last few years, we can count on a relatively rapid DNA technology advance.

Alternative DNA-derived molecules that can store data have also been proposed (Yang, McCloskey, Chaput 2020). These molecules are said to be more stable than DNA and their features need to be investigated more.

As we can see, the world of DNA as storage medium is bustling and presents real perspectives for digital archiving.

Bibliography

ANDERSON, Richard, 2013. The Moab Design for Digital Object Versioning. *The Code4Lib Journal* [online]. 15 July 2013. No. 21. [Accessed 24 June 2021]. Retrieved from: https://journal.code4lib.org/articles/8482

ANTKOWIAK, Philipp L., LIETARD, Jory, DARESTANI, Mohammad Zalbagi, SOMOZA, Mark M., STARK, Wendelin J., HECKEL, Reinhard and GRASS, Robert N., 2020. Low cost DNA data storage using photolithographic synthesis and advanced information reconstruction and error correction. *Nature Communications*. December 2020. Vol. 11, no. 1, p. 5345. DOI 10.1038/s41467-020-19148-3.

ARNOLD, DENIS, FISSENI, BERNHARD, HELFER, FELIX, BUDDENBOHM, STEFAN and KIRALY, PETER, 2020. *Repository Solutions - Technology Watch Report 1* [online]. Zenodo. [Accessed 22 June 2021]. Retrieved from: https://zenodo.org/record/3873027

AUDENHOVE, Leo Van, 2011. Expert Interviews and Interview Techniques for Policy Analysis [document PDF]. [Accessed 15 June 2021]. Retrieved from: https://www.ies.be/files/060313%20Interviews_VanAudenhove.pdf

BALL, Alex, 2006. Briefing Paper--the OAIS Reference Model. [online]. [Accessed 13 July 2021]. Retrieved from: https://www.academia.edu/1145153/Briefing_Paper_the_OAIS_Reference_Model

BANAL, James L., SHEPHERD, Tyson R., BERLEANT, Joseph, HUANG, Hellen, REYES, Miguel, ACKERMAN, Cheri M., BLAINEY, Paul C. and BATHE, Mark, 2021. Random access DNA memory using Boolean search in an archival file storage system. *Nature Materials* [online]. 10 June 2021. [Accessed 14 July 2021]. DOI 10.1038/s41563-021-01021-3. Retrieved from: http://www.nature.com/articles/s41563-021-01021-3

BANCROFT, Carter, BOWLER, Timothy, BLOOM, Brian and CLELLAND, Catherine Taylor, 2001. Long-Term Storage of Information in DNA. *Science*. 7 September 2001. Vol. 293, no. 5536, p. 1763–1765. DOI 10.1126/science.293.5536.1763c.

BERG, Jeremy M., TYMOCZKO, John L., GATTO, Gregory J. and STRYER, Lubert, 2019. *Biochemistry*. Ninth edition. New York: Macmillan International, Higher Eduction. ISBN 978-1-319-11465-7.

BLAWAT, Meinolf, GAEDKE, Klaus, HÜTTER, Ingo, CHEN, Xiao-Ming, TURCZYK, Brian, INVERSO, Samuel, PRUITT, Benjamin W. and CHURCH, George M., 2016. Forward Error Correction for DNA Data Storage. *Procedia Computer Science*. 2016. Vol. 80, p. 1011–1022. DOI 10.1016/j.procs.2016.05.398.

BOGNER, Alexander, LITTIG, Beate, MENZ, Wolfgang, and PALGRAVE CONNECT (ONLINE SERVICE), 2009. *Interviewing experts* [online]. Basingstoke [u.a.: Palgrave Macmillan. [Accessed 13 July 2021]. ISBN 9786612556630.

BORNHOLT, James, LOPEZ, Randolph, CARMEAN, Douglas M., CEZE, Luis, SEELIG, Georg and STRAUSS, Karin, 2016. A DNA-Based Archival Storage System. In: *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems* [online]. Atlanta Georgia USA: ACM. 25 March 2016. p. 637–649. [Accessed 26 June 2021]. ISBN 978-1-4503-4091-5. Retrieved from: https://dl.acm.org/doi/10.1145/2872362.2872397

CALLADINE, C. R., 2004. *Understanding DNA: the molecule & how it works* [online]. 3rd ed. San Diego, CA: Elsevier Academic Press. [Accessed 27 June 2021]. ISBN 9786610968251. Retrieved from: http://www.myilibrary.com?id=96825

CCSDS (The Consultative Committee for Space Data Systems), 2019. *Reference Model for an Open Archival Information System (OAIS) – Pink (pre-magenta) Book* [online]. 2019. [Accessed 9 February 2021]. Retrieved from: https://cwe.ccsds.org/moims/_layouts/15/WopiFrame.aspx?sourcedoc={61C755A7-2C54-4D0D-A8F0-7B6A4228D74C}&file=OAIS%20final%20v3%20draft%20with%20changes%20wrt%20OAISv2%2020190924-rl.docx&action=default

CEZE, Luis, NIVALA, Jeff and STRAUSS, Karin, 2019. Molecular digital data storage using DNA. *Nature Reviews Genetics*. August 2019. Vol. 20, no. 8, p. 456–466. DOI 10.1038/s41576-019-0125-3.

CHANDRA, Smita and GOKHALE, Pratibha, 2012. Implementing Open Archival Information System Model for Digital Preservation at Indian Institute of Geomagnetism. *DESIDOC Journal of Library & Information Technology*. 18 July 2012. Vol. 32, no. 4, p. 327-334.

CHURCH, G. M., GAO, Y. and KOSURI, S., 2012. Next-Generation Digital Information Storage in DNA. *Science*. 28 September 2012. Vol. 337, no. 6102, p. 1628–1628. DOI 10.1126/science.1226355.

CLELLAND, Catherine Taylor, RISCA, Viviana and BANCROFT, Carter, 1999. Hiding messages in DNA microdots. *Nature*. June 1999. Vol. 399, no. 6736, p. 533–534. DOI 10.1038/21092.

DAVIS, Joe, 1996. Microvenus. *Art Journal*. 1996. Vol. 55, no. 1, p. 70–74. DOI 10.2307/777811.

DONG, Yiming, SUN, Fajia, PING, Zhi, OUYANG, Qi and QIAN, Long, 2020. DNA storage: research landscape and future prospects. *National Science Review*. 1 June 2020. Vol. 7, no. 6, p. 1092–1107. DOI 10.1093/nsr/nwaa007.

EISENSTEIN, Michael, 2020. Enzymatic DNA synthesis enters new phase. *Nature Biotechnology*. 1 October 2020. Vol. 38, no. 10, p. 1113–1115. DOI 10.1038/s41587-020-0695-9.

ERLICH, Yaniv and ZIELINSKI, Dina, 2017. DNA Fountain enables a robust and efficient storage architecture. *Science*. 3 March 2017. Vol. 355, no. 6328, p. 950–954. DOI 10.1126/science.aaj2038.

GODDARD, Image credit: Xiaoli Sun, NASA, 2013. *English: To clean up transmission errors introduced by Earth's atmosphere (left), Goddard scientists applied Reed-Solomon error correction (right), which is commonly used in CDs and DVDs. Typical errors include missing pixels (white) and false signals (black). The white stripe indicates a brief period when transmission was paused. Image credit: Xiaoli Sun, NASA Goddard* [online]. 8 January 2013. [Accessed 2 July 2021]. Retrieved from: https://commons.wikimedia.org/wiki/File:Reed%E2%80%93Solomon_error_correction_Mona_Lisa_LroLrLasercomFig4.jpg

GOLDMAN, Nick, BERTONE, Paul, CHEN, Siyuan, DESSIMOZ, Christophe, LEPROUST, Emily M., SIPOS, Botond and BIRNEY, Ewan, 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*. 7 February 2013. Vol. 494, no. 7435, p. 77–80. DOI 10.1038/nature11875.

GRASS, Robert N., HECKEL, Reinhard, PUDDU, Michela, PAUNESCU, Daniela and STARK, Wendelin J., 2015. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition*. 16 February 2015. Vol. 54, no. 8, p. 2552–2555. DOI 10.1002/anie.201411378.

GRIFFITH, Arran, 2021. All Aboard for Fedora 6.0. *Duraspace.org* [online]. 11 June 2021. [Accessed 25 June 2021]. Retrieved from: https://duraspace.org/all-aboard-for-fedora-6-0/

HANKINSON, Andrew, BROWER, Donald, JEFFERIES, Neil, METZ, Rosalyn, MORLEY, Julian, WARNER, Simeon and WOODS, Andrew, 2019. The Oxford Common File Layout: A Common Approach to Digital Preservation. *Publications*. 4 June 2019. Vol. 7, no. 2, p. 39. DOI 10.3390/publications7020039.

HANKINSON, Andrew, JEFFERIES, Neil, METZ, Rosalyn, MORLEY, Julian, WARNER, Simeon and WOODS, Andrew, 2020. *Oxford Common File Layout Specification* [online]. 2020. [Accessed 9 February 2021]. Retrieved from: https://ocfl.io/draft/spec/

HUC, Claude, [no date]. ARCHIVAGE NUMÉRIQUE. *Encyclopædia Universalis* [online]. no date. [Accessed 22 June 2021]. Retrieved from: https://www.universalis.fr/encyclopedie/archivage-numerique/

HUGHES, Randall A. and ELLINGTON, Andrew D., 2017. Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harbor Perspectives in Biology*. January 2017. Vol. 9, no. 1, p. a023812. DOI 10.1101/cshperspect.a023812.

INTERPARES (International Research on Permanent Authentic Records in Electronic Systems), [no date]. The InterPARES 2 Project Glossary. *Interpares.org* [online]. [Accessed 28 August 2021]. Retrieved from: http://www.interpares.org/ip2/display_file.cfm?doc=ip2_glossary.pdf&CFID=26369110&CFTOKEN=57288164

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2012. *Space data and information transfer systems – Open archival information system (OAIS) – Reference model*. 2nd ed. Geneva: ISO, 1st September 2012. ISO 14721.

JEFFERIES, Neil, BREDENBERG, Karin and DAPPERT, Angela, 2019. Aligning the eARK4All Archival Information Package and Oxford Common File Layout Specifications. [online]. 2019. [Accessed 24 June 2021]. DOI 10.17605/OSF.IO/B8G5X. Retrieved from: https://osf.io/b8g5x/

JORDAN, Mark and BARNES, Marcus Emmanuel, 2019. Islandora 8 and Beyond. [online]. 17 June 2019. [Accessed 25 June 2021]. Retrieved from: https://tspace.library.utoronto.ca/handle/1807/95485

KOCH, Traugott and SØLVBERG, Ingeborg Torvik (eds.), 2003. *Research and Advanced Technology for Digital Libraries: 7th European Conference, ECDL 2003 Trondheim, Norway, August 17-22, 2003 Proceedings* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg. [Accessed 22 June 2021]. Lecture Notes in Computer Science. ISBN 978-3-540-40726-3. Retrieved from: http://link.springer.com/10.1007/b11967

KOHLL, A. Xavier, ANTKOWIAK, Philipp L., CHEN, Weida D., NGUYEN, Bichlien H., STARK, Wendelin J., CEZE, Luis, STRAUSS, Karin and GRASS, Robert N., 2020. Stabilizing synthetic DNA for long-term data storage with earth alkaline salts. *Chemical Communications*. 26 March 2020. Vol. 56, no. 25, p. 3613–3616. DOI 10.1039/D0CC00222D.

KUMMER, H., 1987. The consultative committee for space data systems (CCSDS) planned and potential use of the recommendations. *Acta Astronautica*. 1 January 1987. Vol. 16, p. 199–205. DOI 10.1016/0094-5765(87)90106-8.

KUNZE, J., LITTMAN, J., MADDEN, E., SCANCELLA, J. and ADAMS, C., 2018. *The BagIt File Packaging Format (V1.0)* [online]. RFC8493. RFC Editor. [Accessed 24 June 2021]. Retrieved from: https://www.rfc-editor.org/info/rfc8493

LAVOIE, Brian F., 2004. The Open Archival Information System Reference Model: Introductory Guide. *Microform & Imaging Review* [online]. January 2004. Vol. 33, no. 2. [Accessed 22 June 2021]. DOI 10.1515/MFIR.2004.68. Retrieved from: https://www.degruyter.com/document/doi/10.1515/MFIR.2004.68/html

LEE, Henry H., KALHOR, Reza, GOELA, Naveen, BOLOT, Jean and CHURCH, George M., 2019. Terminator-free template-independent enzymatic DNA synthesis for digital information storage. *Nature Communications*. December 2019. Vol. 10, no. 1, p. 2383. DOI 10.1038/s41467-019-10258-1.

LOCKSS, [no date]. Industry Standards | LOCKSS. [online]. [n.d.]. [Accessed 13 July 2021]. Retrieved from: https://www.lockss.org/use-lockss/industry-standards

LOGGAWIGGLER, 2008. Free Image on Pixabay - Sandstone, Landmark, Natural Arch. *pixabay.com* [online]. 2008. [Accessed 25 June 2021]. Retrieved from: https://pixabay.com/photos/sandstone-landmark-natural-arch-4625/

LUNA, Leslie, 2019. UC San Diego Library Receives Mellon Grant to Develop Approaches to Preserving Digital Repositories. *UC San Diego News Center* [online]. 5 February 2019. [Accessed 24 June 2021]. Retrieved from: https://ucsdnews.ucsd.edu/pressrelease/uc_san_diego_library_receives_mellon_grant_to_develop_approaches_to_preserving_digital_repositories

LYNCH, Michael, 2019. FAIR Simple Scalable Static Research Data Repository. *UTS eResearch* [online]. 5 November 2019. [Accessed 25 June 2021]. Retrieved from: https://eresearch.uts.edu.au/2019/11/05/FAIR Repo - eResearch Presentation.htm

LYNCH, Mike, 2019. Publishing versioned datasets using OCFL and nginx. *UTS eResearch* [online]. 5 November 2019. [Accessed 25 June 2021]. Retrieved from: https://eresearch.uts.edu.au/2019/11/05/eResearch2019_lighting_ocfl_nginx.htm

MDN, (Mozilla Developper Network), [no date]. MDN Web Docs Glossary: Definitions of Web-related terms. *https://developer.mozilla.org* [online]. [n.d.]. [Viewed 30 September 2021]. Available from: https://developer.mozilla.org/en-US/docs/Glossary/CRUD

MEISER, Linda C., ANTKOWIAK, Philipp L., KOCH, Julian, CHEN, Weida D., KOHLL, A. Xavier, STARK, Wendelin J., HECKEL, Reinhard and GRASS, Robert N., 2020. Reading and writing digital data in DNA. *Nature Protocols*. January 2020. Vol. 15, no. 1, p. 86–101. DOI 10.1038/s41596-019-0244-5.

MEUSER, Michael and NAGEL, Ulrike, 2009. The Expert Interview and Changes in Knowledge Production. In: BOGNER, Alexander, LITTIG, Beate and MENZ, Wolfgang (eds.), *Interviewing Experts* [online]. London: Palgrave Macmillan UK. p. 17–42. Research Methods Series. [Accessed 13 July 2021]. ISBN 978-0-230-24427-6. Retrieved from: https://doi.org/10.1057/9780230244276_2

MUET, Florence, 2003. *Conduire un entretien semi-directif* [document PDF]. Document interne à la Haute école de gestion de Genève, filière information documentaire.

NDIIPP, National Digital Information Infrastructure and Preservation Program, 2002. *Plan for the National Digital Information Infrastructure and Preservation Program - Part 2, the Appendices* [online]. 2002. [Accessed 22 June 2021]. Retrieved from: https://www.digitalpreservation.gov/multimedia/documents/ndiipp_appendix.pdf

NICHOLSON, Dennis and DOBREVA, Milena, 2009. Beyond OAIS: Towards a reliable and consistent digital preservation implementation framework. In: *2009 16th International Conference on Digital Signal Processing* [online]. Santorini, Greece: IEEE. July 2009. p. 1–8. [Accessed 27 June 2021]. ISBN 978-1-4244-3297-4. Retrieved from: http://ieeexplore.ieee.org/document/5201126/

OASIS, Organization for the Advancement of Structured Information Standards, [n.d.]. OASIS SOA Reference Model (SOA-RM) TC. [online]. [n.d.]. [Accessed 22 June 2021]. Retrieved from: https://www.oasis-open.org/committees/soa-rm/faq.php

ORGANICK, Lee, ANG, Siena Dumas, CHEN, Yuan-Jyue, LOPEZ, Randolph, YEKHANIN, Sergey, MAKARYCHEV, Konstantin, RACZ, Miklos Z, KAMATH, Govinda, GOPALAN, Parikshit, NGUYEN, Bichlien, TAKAHASHI, Christopher N, NEWMAN, Sharon, PARKER, Hsing-Yeh, RASHTCHIAN, Cyrus, STEWART, Kendall, GUPTA, Gagan, CARLSON, Robert, MULLIGAN, John, CARMEAN, Douglas, SEELIG, Georg, CEZE, Luis and STRAUSS, Karin, 2018. Random access in large-scale DNA data storage. *Nature Biotechnology*. March 2018. Vol. 36, no. 3, p. 242–248. DOI 10.1038/nbt.4079.

ORGANICK, Lee, NGUYEN, Bichlien H., MCAMIS, Rachel, CHEN, Weida D., KOHLL, A. Xavier, ANG, Siena Dumas, GRASS, Robert N., CEZE, Luis and STRAUSS, Karin, 2021. An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage. *Small Methods*. May 2021. Vol. 5, no. 5, p. 2001094. DOI 10.1002/smtd.202001094.

PANDA, Darshan, MOLLA, Kutubuddin Ali, BAIG, Mirza Jainul, SWAIN, Alaka, BEHERA, Deeptirekha and DASH, Manaswini, 2018. DNA as a digital information storage device: hope or hype? *3 Biotech*. May 2018. Vol. 8, no. 5, p. 239. DOI 10.1007/s13205-018-1246-7.

PAUNESCU, Daniela, PUDDU, Michela, SOELLNER, Justus O. B., STOESSEL, Philipp R. and GRASS, Robert N., 2013. Reversible DNA encapsulation in silica to produce ROS-resistant and heat-resistant synthetic DNA “fossils.” *Nature Protocols*. December 2013. Vol. 8, no. 12, p. 2440–2448. DOI 10.1038/nprot.2013.154.

PING, Zhi, ZHANG, Haoling, CHEN, Shihong, ZHUANG, Qianlong, ZHU, Sha Joe and SHEN, Yue, 2020. Chamaeleo: a robust library for DNA storage coding schemes. *bioRxiv*. 19 March 2020. P. 2020.01.02.892588. DOI 10.1101/2020.01.02.892588.

ROSENFIELD, Israel, ZIFF, Edward and VAN LOON, Borin, 2011. *DNA: a graphic guide to the molecule that shook the world*. New York: Columbia University Press. ISBN 978-0-231-14270-0.

RUTAKUMWA, Rwamahe, MUGISHA, Joseph Okello, BERNAYS, Sarah, KABUNGA, Elizabeth, TUMWEKWASE, Grace, MBONYE, Martin and SEELEY, Janet, 2020. Conducting in-depth interviews with and without voice recorders: a comparative analysis. *Qualitative Research*. October 2020. Vol. 20, no. 5, p. 565–581. DOI 10.1177/1468794119884806.

SCHUMANN, Natascha and RECKER, Astrid, 2013. De-mystifying OAIS compliance: Benefits and challenges of mapping the OAIS reference model to the GESIS Data Archive. *IASSIST Quarterly*. 6 November 2013. Vol. 36, no. 2, p. 6. DOI 10.29173/iq769.

SEFTON, Peter, 2019. Implementation of a Research Data Repository using the Oxford Common File Layout standard at the University of Technology Sydney. *UTS eResearch*

[online]. 1 July 2019. [Accessed 25 June 2021]. Retrieved from: https://eresearch.uts.edu.au/2019/07/01/OCLF.htm

SEFTON, Peter, 2020. An open, composable standards–based research eResearch platform: Arkisto. *UTS eResearch* [online]. 23 November 2020. [Accessed 25 June 2021]. Retrieved from: https://eresearch.uts.edu.au/2020/11/23/Arkisto.htm

SHENDURE, Jay and JI, Hanlee, 2008. Next-generation DNA sequencing. *Nature Biotechnology*. October 2008. Vol. 26, no. 10, p. 1135–1145. DOI 10.1038/nbt1486.

STANLEY, Philip M., STRITTMATTER, Lisa M., VICKERS, Alice M. and LEE, Kevin C.K., 2020. Decoding DNA data storage for investment. *Biotechnology Advances*. December 2020. Vol. 45, p. 107639. DOI 10.1016/j.biotechadv.2020.107639.

THIBODEAU, Kenneth, [no date]. Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years • CLIR. *CLIR* [online]. [n.d.]. [Accessed 14 July 2021]. Retrieved from: https://www.clir.org/pubs/reports/pub107/thibodeau/

THOMAS, Oliver, 2006. Understanding the Term Reference Model in Information Systems Research: History, Literature Analysis and Explanation. In: BUSSLER, Christoph J. and HALLER, Armin (eds.), *Business Process Management Workshops* [online]. Berlin, Heidelberg: Springer Berlin Heidelberg. p. 484–496. Lecture Notes in Computer Science. [Accessed 22 June 2021]. ISBN 978-3-540-32595-6. Retrieved from: http://link.springer.com/10.1007/11678564_45

WEITZMAN, Jonathan and WEITZMAN, Matthew, 2020. *30-Second Genetics The 50 Most Revolutionary Discoveries in Genetics, Each Explained in Half a Minute.* [online]. Minneapolis: Ivy Press, The. [Accessed 27 June 2021]. ISBN 978-1-78240-619-8. Retrieved from: http://public.eblib.com/choice/PublicFullRecord.aspx?p=6181415

WIB, (What is Biotechnology), [no date]. WhatisBiotechnology • The sciences, places and people that have created biotechnology. *WhatisBiotechnology.org* [online]. [s.d.]. [Accessed 26 June 2021]. Retrieved from: https://www.whatisbiotechnology.org/

WILSON, Thomas C., 2017. Rethinking digital preservation: definitions, models, and requirements. *Digital Library Perspectives*. 8 May 2017. Vol. 33, no. 2, p. 128–136. DOI 10.1108/DLP-08-2016-0029.

YANG, Kefan, MCCLOSKEY, Cailen M. and CHAPUT, John C., 2020. Reading and Writing Digital Information in TNA. *ACS Synthetic Biology*. 20 November 2020. Vol. 9, no. 11, p. 2936–2942. DOI 10.1021/acssynbio.0c00361.

Appendix 1: Interview report – Common part

All three first series of interviews were held in French. Consequently, the reports all had a common part that described the process behind the project and the interview itself. This text can be found in the following section. The second appendix will set forth the first expert opinion, the third appendix the second expert's and the appendix 3 will contain the conclusion from interview number 3. For anonymity concern, the name and experts' profiles were removed from initial reports.

Résumé analytique

Le projet « OAIS compliant digital archiving in DNA » est né dans le cadre d'un travail de bachelor en Information Documentaire au sein de la Haute École de Gestion de Genève (HEG-GE). Ce projet de fin d'études vise à une modélisation d'un système d'archivage des données numériques, conforme au modèle de référence Open Archival Information System (OAIS), utilisant l'acide désoxyribonucléique (ADN) comme support de stockage.

L'ADN est le support de l'information génétique, garant de la pérennité du vivant depuis l'apparition de la vie. L'idée d'utiliser l'ADN comme support d'information n'est pas récente : ses principales caractéristiques en font un candidat idéal du fait de la densité d'information qu'il peut contenir, de sa relative résistance mais surtout son statut au fondement du vivant le préservant de toute obsolescence technologique.

Les questions de la pérennité et de la volumétrie ont de tout temps préoccupé les archivistes. En effet, la durée de vie d'un medium de stockage n'est pas infinie et de même, l'avancée technologique entraîne une rapide incompatibilité des supports (dans un délai d'environ 5 ans) ce qui impose des migrations régulières des contenus. La problématique du volume d'information auquel les professionnels sont confrontés chaque année reste actuelle et pressante.

Face à ces préoccupations, l'ADN semble apporter une solution dont la concrétisation nécessite d'une part une adéquation à OAIS mais aussi une définition des particularités techniques qui permettront une implémentation efficace et efficiente de ce medium à un système de préservation à long terme.

C'est dans cette optique que cet entretien avec un expert a été réalisé. Il est indispensable d'établir quelles exigences sont nécessaires au système pour remplir les critères d'OAIS et quelles solutions technologiques concrètes peuvent être utilisées pour une appréhension optimale du sujet. La présente entrevue est la première dans une série de trois et à l'issue de chaque entretien, les exigences techniques seront calibrées fournir un modèle un large éventail de champs d'actions pour servir à différents systèmes de stockage. Ces spécifications seront ensuite présentées à un groupe de recherche spécialisé dans les différentes applications de l'ADN.

Introduction

Il s'agit dans cette partie du projet d'aborder un aspect technique qui n'est pas évoqué dans le modèle de référence. Ce dernier qui se veut agnostique décrit l'organisation d'un système

d'archivage numérique de manière théorique, indiquant les fonctions et les contraintes qui engagent le système sans en préciser concrètement les pendants techniques.

L'entretien est effectué à l'issue d'une revue de la littérature sur OAIS et plus particulièrement sur le standard d'architecture de la hiérarchie de stockage Oxford Common File Layout (OCFL), une proposition concrète pour l'agencement et la prise en charge des fichiers numériques dans un système robuste et autosuffisant.

La tenue de cette entrevue avait pour but de clarifier/préciser des idées de spécifications concrètes de telle manière que le modèle puisse être implémenté. D'autre part, l'utilisation de l'ADN implique une couche de technicité qui exige une définition claire préalable de ce que l'on cherche à faire.

Méthodologie

Activités préparatoires de l'entretien

Le choix d'une session semi-structurée s'imposait car il s'agit dans un premier temps d'exposer une synthèse de ce qui a été tiré de la revue de la littérature ainsi que des réflexions menées avec le mandant en ce qui concerne les spécifications et dans un deuxième temps, obtenir des experts un retour et générer de nouvelles idées. Pour délimiter et éclaircir le sujet, une présentation a été préparée au préalable afin de diriger la discussion mais surtout susciter des idées et des concepts nouveaux pour que les spécifications soient le plus exhaustif possible.

Sélection du lieu

Tous les participants ont opté pour un déroulement en virtuel de l'entrevue.

Sélection des participants

Les experts ont été ciblés et choisis car ils sont tous issus d'institutions qui ont chacun développé et mis en place leur propre système d'archivage électronique, conforme à la norme ISO 14721. Chacun de ces systèmes constitue une interprétation différente mais correcte de la norme. Ces personnes possèdent donc une connaissance concrète et précise du modèle OAIS ainsi que ses exigences.

D'autre part, ils jouissent tous d'une assise et d'une renommée dans la communauté pour leurs travaux dans le domaine de l'archivage électronique.

Finalement, ce sont des personnes qui ont exprimé leur enthousiasme vis-à-vis du projet et elles étaient sincèrement curieuses de voir l'évolution et l'aboutissement du présent travail.

Appendix 2: Expert 1 point of view

Interview d'expert

L'entrevue s'est déroulée le lundi 15 mars 2021 pendant une session d'une durée de 90 minutes. Outre l'expert, étaient présents Monsieur Jan Krause-Bilvin en tant que modérateur et Dina Andriamahady, instigatrice de l'interview. Le modérateur et l'animatrice ont procédé à des prises de notes tout au long de la séance. Ceci a conduit ensuite à une séance de débriefing et de mise en commun des notes prises.

Discussion des résultats

Réflexions relatives à l'entité stockage de OAIS

La discussion a soulevé la nécessité de prioriser les trois éléments suivants :

- L'identifiant de l'AIP
- Le bitstream
- Les checksums

Les points suivants ont aussi été abordés :

- La hiérarchie de stockage : suggestion d'un stockage directement à la racine du stockage OCFL.
- Les contraintes que peuvent représenter un error checking régulier, pourtant recommandé par OAIS, et la pertinence de plutôt multiplier les copies dans le cas du medium ADN.

Pour ce qui est du choix de formats, l'expert a reconnu que OCFL est la production la plus récente en matière de standards cependant il serait certainement intéressant de considérer d'autres paquets et systèmes de fichiers distribués. Ont été cités notamment : Nodeshape, Memento, Centera ou Ceph.

Réflexions relatives à l'utilisation de l'ADN

Compte tenu de la déficience du procédé en termes de réactivité, l'ADN serait plutôt indiqué comme solution pour le « disaster recovery ». Et même dans ce cas il faut prendre en considération que le temps de récupération doit être adapté au fonctionnement de l'institution. Dans le cas d'archives cantonales, quelques (2-3) mois peuvent être acceptables. S'il y a de nombreuses copies sur un médium stable, tel que l'ADN, des check intégraux ne sont que très rarement nécessaires.

Réflexions générales

Pour ce qui est des considérations techniques et volumétriques, l'expert recommande :

- Un encodage en langage UTF qui serait idéal pour pouvoir rendre tous les caractères et permettre une polyvalence et une exhaustivité du système.
- Une assez grande capacité de stockage : en effet, la volumétrie pour un canton généralement avoisine 1 Pétabyte sans les données audiovisuelles. Il faudrait prévoir alors, dans le court terme, un système pouvant accueillir 10 Pétabytes (sans l'audiovisuel).

À noter aussi :

- L'importance du "versionning" ne serait-ce que dans le cas des métadonnées, notamment dans le cadre des actions de préservation ;
- L'intérêt de la déduplication, acceptable au sein des AIP ;
- La nécessité de garantir l'authenticité, en particulier, la fixité des métadonnées encore peu pratiquée ;
- Réviser les métadonnées afin qu'elles puissent documenter le travail des archivistes.

Conclusion

L'expert a partagé ses perceptions de la synthèse de notre revue systématique et de notre conception des spécifications nécessaires à l'application de l'utilisation de l'ADN comme support de l'information numérique. Outre les propositions et recommandations évoquées plus haut qui nous permettront de compléter notre manifeste de spécifications, cela nous a permis de repenser des détails qui ont fait surface pendant la discussion.

De cette entrevue, nous sommes arrivés aux nouvelles considérations suivantes :

- L'idée d'utiliser les checksums comme identifiants des AIP pourrait être appliquée au stockage ADN

  exemple :
  SHA512([racine_stockage]/[hiérarchie]/[objetOCFL]/[cheminDansObjet]/[fichier])
- Imaginer une formule qui en fonction de la demi-vie de la stabilité de l'information sur l'ADN permet de dire en fonction du pourcentage de copies encore lisible l'intervalle de temps pour laquelle un check complet et une correction sont nécessaires. Si le seuil d'erreur est dépassé, il faudrait resynthétiser l'ADN et remplacer les anciennes copies.

Annexe

⇨ Support de la présentation du 13 mars 2021

Appendix 3 : Expert 2 point of view

Discussion des résultats

Le standard OCFL est récent et encore très peu éprouvé. Il y a de fortes chances que le standard ne soit jamais adopté massivement par la communauté contrairement à BagIt dont l'usage est largement répandu et dont les bibliothèques sont plus mûres.

D'autre part, le versioning proposé par OCFL pose une problématique au niveau archivistique. Il est même antinomique de l'archivage qui est censé concerner des documents scellés.

Une approche est de considérer les AIP comme immuables pour éviter tout risque de confusion lors des citations. C'est l'approche adoptée par OLOS, une solution de préservation axée sur la reproductibilité de la recherche. Le pragmatisme impose de n'avoir qu'une seule « bonne version ». Les relations entre les « versions » peuvent être décrites dans les métadonnées via des références entre AIP.

Conclusion

Les experts ont partagé leurs perceptions de la synthèse de notre revue systématique et de notre conception des spécifications nécessaire à l'application de l'utilisation de l'ADN comme support de l'information numérique. Outre les remarques évoquées plus haut qui nous permettront de compléter notre manifeste de spécifications, cette séance nous a permis de repenser des détails qui ont fait surface pendant la discussion.

De cette entrevue, nous sommes arrivés à la nouvelle considération suivante : il serait recommandé de ne pas se cantonner à ce qui est offert par OCFL mais considérer les avantages de BagIt car le versioning est contraire au principe archivistique qui stipule qu'on ne doit pas modifier un AIP.

Annexe

- ⇨ Support de la présentation du 26 mars 2021

Appendix 4 : Expert 3 point of view

Discussion des résultats

Réflexions relatives à l'entité stockage de OAIS

OAIS prévoit la mise à jour des AIP. Dans la pratique cela revient à gérer les versions des données au sein des AIP ce qui implique :

- La nécessité de se prémunir contre l'effacement des données
- L'importance de conserver les anciennes versions
- L'obligation d'identifier de manière sans équivoque les différentes versions entre elles (pour référence ou lors de citation directe)

Au CERN, chaque fichier constitue un seul AIP et les versions sont gérées dans des AIC (Archival Information Collection) qui sont des collections d'AIP. Les AIP sont immuables et il n'est pas nécessaire de faire des mises à jour sur les inventaires. Ce système basé sur BagIt utilise le timestamp+md5 comme identification et a pour avantage d'éviter d'avoir un poids trop lourd pour un AIP.

Réflexions générales

Il sera indispensable d'inclure une documentation à long terme indépendante de la technologie ADN. Il faudra y inscrire :

- La documentation ad hoc pour les informations concernant l'ensemble du système de préservation
- La structure de la hiérarchie de stockage (algorithme,...)
- Le standard AIP
- Le schéma des métadonnées
- Le choix relatif à l'implémentation du média ADN (même s'il s'agit d'une technologie à très faible risque d'obsolescence).

Conclusion

L'expert a partagé ses perceptions de la synthèse de notre revue systématique et de notre conception des spécifications nécessaire à l'application de l'utilisation de l'ADN comme support de l'information numérique. Outre les propositions et recommandations évoquées plus haut qui nous permettront de compléter notre manifeste de spécifications, cela nous a permis de repenser des détails qui ont fait surface pendant la discussion.

De cette entrevue, nous sommes arrivés aux nouvelles considérations suivantes :

- L'utilisation des checksums des fichiers comme identifiant : cela garantit leur unicité et évite les problèmes de confusion.
- L'importance d'une documentation complète qui sera impérativement à préserver hors du système de stockage.

Annexe

⇨ Support de la présentation du 01 avril 2021

Appendix 5: DNA expert interview report

Summary of project

The project "OAIS compliant digital archiving in DNA" was born within the framework of a bachelor's degree in Library and Information Science at the Haute École de Gestion de Genève (HEG-GE), member of the University of Applied Sciences Western Switzerland (HES-SO). This end-of-studies project aims at modeling a digital data archiving system, compliant with the Open Archival Information System (OAIS) reference model, using deoxyribonucleic acid (DNA) as storage media.

DNA carries the genetic information of living beings: it contains the information necessary for the development and maintenance of living organisms. The idea of using DNA as digital information carrier has been around for quite some time but lack of technology prevented its completion. Its main characteristics make it an ideal candidate because of the density of information it can contain, its relative longevity, but above all its status at the foundation of living beings, preserving it from any technological obsolescence.

Long-term preservation and data volume issues have burdened archivists for a long time. Indeed, the lifespan of an average storage media is rather short in archival perspective and similarly, technological progress leads to a rapid incompatibility of media (within a period of about 5 years), which requires regular migrations of content. The sheer volume of information that professionals have to handle each year remains a current and pressing matter.

Faced with these concerns, DNA seems to provide a solution whose realization requires first conformance to the Open Archival Information System (OAIS), a reference model, which is an authority in the world of libraries, archives and museums. Its enforcement also necessitates a definition of the technical characteristics that will allow an effective and efficient implementation of this media to a long-term preservation system.

It is with this in mind that interviews with experts were held. Exchanging with experts in the OAIS field was a prerequisite to establish which requirements are necessary for the system to fulfill the OAIS criteria and which effective technological solutions to use for an optimal fulfillment of the model.

Ensuing these first series of interviews, it was finally time to set up a meeting with the “DNA people” in order to determine to what extent those expectations were doable. This very report summarizes the outcome of these exchanges.

Introduction

In this part of the project, we address a technical aspect that the OAIS reference model did not broach. The latter, which is intended to be agnostic, describes the organization of a digital archiving system in a theoretical manner, indicating the functions and constraints the system has to fulfill without specifying any technical counterparts.

We conducted a review of the literature on OAIS and more specifically on the Oxford Common File Layout (OCFL), a storage hierarchy standard that brings a practical proposal for the

arrangement and support of digital files in a robust and self-sufficient system. The defined specifications were then presented to OAIS experts to ascertain their completeness.

The purpose of holding this very interview was to bring together all those specifications and confront them to the experimental DNA field to see how they could be implemented.

Methodology & Participant profile

Instrument development

The choice of a semi-structured session was necessary. To delimit and clarify the subject, a presentation was prepared beforehand in order to direct the discussion but especially to present a synthesis of what was drawn from the literature review and the reflections carried out regarding the specifications.

Site selection

All participants opted for a virtual interview process.

Participant selection

We targeted a research group that has been working on using DNA as media for digital information since 2015 at ETH Zürich (ETHZ). Their research aims to make DNA an alternative device for long-term preservation of data.

Expert interview session

The interview took place on Friday, May 28, 2021 during a 90-minute session. In addition to the expert interviewee, Mr. Jan Krause-Bilvin was present as the moderator and Dina Andriamahady, the initiator of the interview. The moderator and the facilitator took notes throughout the session. This led to a debriefing session and sharing of notes.

Discussion results

Storage hierarchy feasibility

Encoding the file path and the filename (unique identity) in the primer is doable. They would be then encoded as binary files thus encompassing the use of Unicode for file names. Using the primer sequence as the identifier would allow random file pick up regardless of file layout. A difference in the “depth” of the deigned primers would allow determinant retrieval whether only a file is needed or the whole package (AIP) is required.

However, a primers library would have to be kept separately as a written record. If primers list happens to be lost or in a worst-case scenario, the whole library could then be obtained back after sequencing the data.

If the AIP are zipped up into “.tar” files, for instance, only the entirety of the AIP could be extracted.

Data capacity

The theoretical limit for DNA storage is 17 Exabytes per gram (at 10 copies physical redundancy) but there are various limitations. For instance, synthesis is certainly bottleneck

(mainly at costs level) but there are other parameters such as redundancy (logical and physical) to take into account.

Metadata are to be taken care of as part of the data at the root of the content.

Versioning

With different series of designed primers, versioning could be achieved since they would guarantee that different versions could be discriminated against each other. Encapsulated DNA could be added to a tube already containing the previous version.

However, file updates such as is recommended in the OCFL standard is inefficient and should be avoided.

Storage

Silicate beads encapsulation is the preferred route for the Grass research group since the technology is available (developed by Pr. Grass himself). It is a relatively strong physical protection, as bone would be to DNA in fossils, and can guarantee its longevity. Hydrofluoric acid in small concentration is needed to extract DNA from its glass armor ensuring water resistance.

Latency

Outsourcing DNA synthesis in more practical, an estimated speed of 1 to 5 minutes per base is a possible expectation however subcontracting means having little impact on the workflow. Even though a unique stand can be synthesized in each well of the micro-array card, the number of strands really depends on the file size.

The most optimistic reading speed is estimated at 300 kilobytes per second without any redundancy and with the most performant available (benchtop?) sequencer (about 520 000 dollars).

Encoding and correcting algorithms

Reed-Solomon is a tried and tested code in the telecommunication industry. The algorithm is stable; the correction rate is significantly close to that of Fountain code. The same goes for the net information density of the two codes (1.78 bits per nucleotide versus 1.98 bits per nucleotide respectively).

Reed-Solomon entails more logical redundancy which makes it more expensive (adds more length that needs to be synthesized).

Costs

Costs are still high. To avoid transfer of information, buying a sequencer would be a good thing. Even though it is a one-time deal, it is still costly.

DNA synthesizing manufacturers have a price range of about 300$ for 200 kilobases.

DNA manipulations (PCR, sequencing, en/decapsulation) require manual handling which means it is essential to hire at least a laboratory technician.

Conclusion

This exchange with Ms. Meiser was tremendously insightful and allowed us to determine what could be done with DNA technology, how it should be done and what are the perspectives we should look forward to in order to make it practical to use DNA as media for digital storage. Thanks to the information gathered, we could come up with a complete scenario of how to complete an experiment on digital information preservation in DNA held in an archival institution.

Appendix

- ⇨ Presentation support used on 2021 May 26th.