Mémoire / Septembre 2024

Diplôme national de master

Domaine - sciences humaines et sociales

Mention – sciences de l'information et des bibliothèques

Parcours – Archives numériques

MEMOIRE

Etude de faisabilité de stockage sur ADN

NEHDI Kawther

Sous la direction de Lamia BADRA
Maître de conférences en sciences de l'information et de la communication – Laboratoire de recherche ComSocs et Département Métiers de la culture

Remerciements

Je dédie ce travail à ma directrice de mémoire, Madame BADRA Lamia, c'était un honneur de travailler sous votre direction. Je vous exprime ma gratitude la plus sincère pour vos remarques et vos conseils.

Je tiens à exprimer ma profonde gratitude envers le président du jury, Monsieur DUPLOUY Laurent pour l'intérêt qu'il a porté à mon travail, et pour avoir accepté de le juger.

A mes chers parents, mes sœurs, mon frère et ma deuxième famille en France qui ont toujours été présents pour moi. Je vous remercie pour ce que vous avez fait.

Mes sincères remerciements à tous les professeurs, mes amis et à tous ceux qui ont accepté de me rencontrer et de répondre à mes questions au cours de ma formation et de la préparation de ce mémoire.

Résumé :

Parallèlement au développement technologique, la quantité des données numériques augmente. Cependant, ces données doivent être conservées dans les meilleures conditions et sur des supports de stockage fiables afin de garantir dans le temps leurs intégrité, sécurité et accessibilité. Ce mémoire a pour but de présenter les obligations relatives à la bonne conservation des données numériques ainsi que l'étude de faisabilité du stockage sur les supports d'ADN qui est en cours de développement dans le but d'éliminer ses problématiques pour qu'il soit le nouveau support fiable de futur.

Descripteurs :

Archivage, stockage, donnée, information, support, support de stockage, numérique, ADN,

Abstract :

In parallel with technological development, the quantity of digital data is increasing. However, this data must be preserved in the best possible conditions and on reliable storage media in order to guarantee its integrity, security and accessibility over time. The aim of this dissertation is to present the conditions under which digital data must be stored, and the feasibility study of storage on DNA media, which is currently being developed with the aim of eliminating its problems and making it the new reliable medium of the future.

Keywords :

Archiving, storage, data, information, medium, storage media, digital, DNA.

Sommaire

Sigles et abréviations

Acide Désoxyribonucléique (ADN)

Adénine (A)

Application Métier (AI)

Bibliothèque Nationale De France (BNF)

Centre De Calcul De L'in2p3 (CC-IN2P3)

Centre Informatique National De l'Enseignement Supérieur (CINES)

Centre National De La Recherche Scientifique (CNRS)

Cytosine (C)

European Bioinformatics Institute (EBI)

Flexible Image Transport System (FITS)

Guanine (G)

Institut Nationale De l'Audiovisuel (INA)

Institute Of Electrical And Electronics Engineers (IEEE)

Linear Tape-Open (LTO)

Next Generation Sequencing (NGS)

Réaction En Chaine Par Polymérase (PCR)

Système d'Archivage Électronique (SAE)

Système d'Information (SI)

Système De Gestion Electroniques Des Document (GED)

Thymine (T)

INTRODUCTION

Les Calculi[1], les palettes de pierre, les os des animaux, le papyrus, le parchemin et le papier : soit l'ensemble des supports qui ont été utilisés de 30 000 à 25 000 avant notre ère pour sauvegarder et transmettre les informations jusqu'à l'invention de l'écriture et de l'imprimerie. Ces derniers ont servi aussi à accélérer la propagation des informations autour du monde. De même avec l'invention d'autres supports tels que les microfilms, les cartes perforées, et également l'avènement de l'ère numérique qui a donné naissance à d'autres générations de supports.

En effet, l'objectif des spécialistes de l'information est de préserver et de pérenniser les informations qui ne cessent de s'accroître. Réussissent-ils à assurer ces obligations face aux problèmes de ces supports éphémères et obsolètes ? Comment garantir dans le temps et sans modification possible l'intégrité, l'authenticité, la fiabilité et l'accès aux données numériques ?

En effet, une vidéo qui a été publiée sur YouTube nous a totalement captivés il y a deux ans. Elle aborde un sujet très important, qui intéresse absolument les archivistes, en s'interrogeant sur la fiabilité des supports sur lesquels les données sont stockées aujourd'hui dans le monde. Ces données présentent un élément crucial pour l'intelligence artificielle, l'évolution de l'éducation et de la technologie. En effet, elles continuent à augmenter parallèlement avec l'évolution d'internet, or « 1 octet est composé d'une alternance de huit 0 et 1, ainsi l'affichage d'une lettre correspond à un octet, une page 3Ko, 300 pages 1 Mo, une bibliothèque 1Go, cinq bibliothèques un DVD, 6 millions de livres 1To ou une pile de 100 DVD, une pile de DVD haute de 200 mètres représente 1Po, une pile de DVD de 1Km présente 5Eo, soit toute l'information produite par l'humanité jusqu'à 2003. Une pile de DVD reliant la terre à la lune représente 1.8Zo, soit l'information produite durant l'année 2011. Une pile de DVD reliant mars et soleil représente 1Yo, soit le volume de l'information numérique généré de 2022 à 2027. »[2]

[1] Jetons d'argile

[2] SLICE Peuples. (2021, mars 5). *L'ADN, un média de stockage pour le « Big Data » ? SLICE* [Vidéo]. YouTube. Consulté le 2 avril 2024, à l'adresse https://www.youtube.com/watch?v=Ev0Z7d3f4UY

En conséquence, jusqu'à nos jours, il ne figure pas dans le temps des supports de stockage stables et durables pour assurer la pérennisation d'une telle masse de données pour une longue durée sans altérer leurs contenus.

De surcroît, le biologiste Nick Goldman et son équipe de *European Bioinformatics Institute* (*EBI*) en Angleterre ont exploré une nouvelle technique de stockage de données. Il s'agit de l'Acide DésoxyriboNucléique (ADN), présent depuis l'apparition de l'humanité sur terre, donc il s'agit d'un moyen de stockage permanent et stable. En effet, Nick Goldman a indiqué qu'il y a tous les deux ou trois ans une production d'une nouvelle machine capable de créer et de déchiffrer l'ADN puisqu'il s'agit toujours d'une partie du génome des êtres vivants[3].

Plusieurs autres biologistes et laboratoires autour du monde ont parlé de cette technologie, et à chaque fois, ils ajoutent des solutions ou d'autres inventions dans le but d'éliminer les failles de cette technologie. Notamment, une collaboration d'EURECOM avec les Archives Nationales Danoises a donné naissance en 2019 à un projet nommé OligoArchive, qui consiste en l'utilisation d'ADN comme une solution aux problèmes liés à la préservation numérique[4]. De plus, le 23 novembre 2021, les deux premières capsules d'ADN ont été déposées aux Archives Nationales de France, plus exactement dans l'armoire de fer. Il s'agit d'un projet impressionnant de BIOMEMORY grâce aux deux biologistes de l'Université de Sorbonne et de Centre national de la recherche scientifique (CNRS), Pierre Crozet et Stéphane Lemaire, qui ont utilisé des ADN biocompatibles pour le stockage de données[5]. Par conséquent, en quoi consiste cette nouvelle technologie et comment pouvons-nous l'exploiter ? Quelles données stocker ? Quels sont les différents enjeux et limites des données stockées sur l'ADN ? Comment accéder à ces données ? Restent-elles toujours intègres et fiables?

---

[3] Goldman, Nick & Bertone, Paul & Chen, Siyuan & Dessimoz, Christophe & Leproust, Emily & Sipos, Botond & Birney, Ewan. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 494. Consulté le 2 avril 2024, à l'adresse https://doi.org/10.1038/nature11875

[4] I'MTech. (2022, 18 mars). *L&rsquo ; ADN pour stocker les données*. I'MTech. Consulté le 2 avril 2024, à l'adresse https://imtech.imt.fr/2020/03/25/ladn-pour-stocker-les-donnees-oligoarchive/

[5] Karayan, R. (2021, 8 décembre). Les Archives Nationales inaugurent le stockage numérique sur ADN. *www.usine-digitale.fr*. Consulté le 2 avril 2024, à l'adresse https://www.usine-digitale.fr/article/les-archives-nationales-inaugurent-le-stockage-numerique-sur-adn.N1162567

Ce mémoire aborde donc un sujet très intéressant dans le domaine des sciences de l'information qui n'a pas étudié dans les travaux précédents.

En outre, le sujet a été discuté des points de vue des informaticiens et des biologistes dans des disciplines différentes telles que l'informatique et les sciences biologiques. En effet, il existe des travaux qui expliquent bien cette nouvelle technologie issue généralement des biologistes ou des informaticiens. Ce sujet va être abordé dans le domaine des sciences de l'information.

Etant donné que ce mémoire intéresse des personnes provenant de différents domaines, il était important de faire une partie consacrée à l'archivage des données numériques ainsi que les supports utilisés dans le monde afin de clarifier l'écosystème des archives numériques. Pour ce faire, nous avons utilisé diverses sources, parmi elles, le mémoire *Histoire des supports de stockage : de la carte perforée à la clé USB* de FLERMOND Richard qui a présenté les différents supports de stockage durant les différentes générations, leurs avantages et leurs inconvénients.

Ensuite, pour expliquer et analyser l'ADN, son utilité et son rôle pour le stockage des données numériques, nous sommes appuyés sur deux sources différentes. La première source est le dossier de presse *DNA Drive La Révolution de l'ADN Première mondiale : dépôt d'archives numériques encodées sur ADN aux Archives nationales* élaboré suite à la coopération de Sorbonne Université, le CNRS, l'Archives Nationales, & Ministère de la Culture en 2021. La deuxième source, le séminaire de Stéphane Lemaire en 2021 : *DNA DRIVE : une nouvelle technologie de stockage numérique durable*. De plus, nous avons adopté la méthodologie de présentation de ce séminaire afin de structurer et d'organiser nos idées dans les différentes parties de ce travail.

Puis, afin de compléter nos recherches avec une analyse pratique, nous avons utilisé une enquête qualitative, qui nous a permis de mener différents entretiens afin d'élargir notre champ de recherches et d'avoir de nouvelles idées. Pour atteindre ces objectifs, nous avons réussi à réaliser des entretiens avec trois personnes provenant de différents domaines : archives et biologie.

Ces entretiens ont été conduits en visioconférence avec deux participants et par téléphone avec un participant. Nous avons utilisé pour chaque entretien une

grille des questions (voir annexe 3, 5 et 7) spécifiques à chaque domaine de personnes interrogées.

En outre, les résultats de ces entretiens, qui englobent les opinions, les avis et les expériences de ces personnes, ont été bien analysés et explorés afin d'être mis en exergue dans les différentes parties du mémoire.

Ainsi, ce mémoire vise à combler un manque de travaux en archivistique, tant du côté des chercheurs et des chercheuses en archivistique que des professionnels et des spécialistes de l'information. Il s'agit d'une problématique très importante puisqu'elle s'intéresse à une des obligations fondamentales des archivistes, à savoir l'obligation d'être en capacité de démontrer, quel que soit le moment, la fiabilité et l'intégrité des données.

Ces recherches nous ont donc conduit à nous interroger sur cette nouvelle technologie et sur son lien avec l'archivage de données, notamment sur la manière d'évaluer la faisabilité de l'archivage numérique sur l'ADN. Nous souhaitons voir à quel point ces nouveaux supports sont appropriés pour pérenniser les archives numériques.

Ce mémoire est structuré en trois chapitres. Le premier chapitre « archivage de données numériques » explore d'une manière générale les techniques liées à la conservation et la préservation des données numériques en mettant en évidence les défis des formats et des supports de stockage. Le deuxième chapitre « le stockage sur l'ADN » décrit la nouvelle technologie de stockage sur les brins d'ADN en précisant son évolution depuis sa création jusqu'à nos jours. Le troisième chapitre « Étude de faisabilité du stockage sur des brins d'ADN » évoque le cas d'ADN Biocompatible de BIOMEMORY en s'appuyant sur les résultats d'enquêtes de terrain.

PARTIE 1 : ARCHIVAGE DES DONNEES NUMERIQUES

I. BREF HISTORIQUE DES SUPPORTS DE STOCKAGE

Les supports de stockage ont permis depuis des années de transférer les données numériques dans le temps. Ces supports matériels de stockage de l'information se divisent en cinq générations principales dont chacune a ses propres caractéristiques, avantages et limites selon les utilisations.

En effet, la première de ces générations, celle dite des supports physiques ou analogiques, concerne des supports apparus depuis le 18e siècle tels que la carte et la bande perforée, qui sont considérées depuis la fin des années 1990 comme des supports obsolètes. Ensuite, la deuxième génération, également appelée celle des supports magnétiques qui sont utilisés comme supports d'archivage des données grâce à leur grande capacité d'enregistrement, leur faible coût et leur facilité d'utilisation. Parmi ces supports, il y a les bandes magnétiques, les disquettes et les disques durs.

La troisième génération est appelée génération des supports optiques numériques qui permettent l'enregistrement de l'audio et des vidéos en utilisant une technique du laser, tels que les disques compacts CD, le DVD et le Blu-ray. En ce qui concerne la quatrième génération, dite la mémoire flash, qui est utilisée généralement avec des appareils de type appareils photos, téléphones portables, *etc.* Appartiennent à cette génération les clés USB en anglais Universal Serial Bus, les cartes SD en anglais Secure Digital et les cartes mémoires en anglais MS Memory Stick.[6]

Enfin, la dernière et cinquième génération est encore en cours de développement dans le but d'éliminer les problématiques liées aux précédentes générations tels que les problèmes d'obsolescence, de densité, de coût et de lieux de conservation de ces

[6] Flermond, F. R. (2017). *Histoire des supports de stockage : de la carte perforée à la clé USB* (ENSSIB, Éd.) [Mémoires Master « Archives numériques », ENSSIB]. Consulté le 6 avril 2024, à l'adresse https://www.enssib.fr/bibliotheque-numerique/documents/67744-histoire-des-supports-de-stockage-de-la-carte-perforee-a-la-cle-usb.pdf

supports. Parmi ces nouveaux supports, il existe le Cloud qui offre diverses fonctionnalités au niveau de l'accessibilité et de la flexibilité pour le stockage des données numériques. Il y a également le stockage moléculaire, c'est-à-dire le stockage sur des capsules d'ADN. Cette nouvelle technologie a commencé à être utilisée et appliquée dans divers domaines depuis 2013.

En effet, chacune de ces générations a constamment essayé de réduire les problématiques du stockage dans le but de rendre les informations mieux accessibles à tout moment en garantissant à long terme leur intégrité, exploitabilité, fiabilité et sécurité.

II. EXPLICATION D'UTILISATION DU TERME DONNEES NUMERIQUES

➢ Les données numériques VS archives numériques ?

Les données numériques désignent l'ensemble des informations stockées ou codées sous une forme numérique, exprimés par des bits 0 et 1 et qui sont exploitables par les machines. Aussi, ces données numériques comprennent davantage des formes d'informations que les archives physiques. Ces données peuvent prendre la forme des vidéos, images, textes et documents sonores.

En effet, ces données sont générées rapidement avec le développement du numérique, dans des domaines (la médecine et l'économie.), et des sources (les réseaux sociaux et les bases de données gouvernementales, *etc.*)[7]

En outre, ces données ont une valeur qui rend leur archivage obligatoire. Dès lors, on peut se demander s'il faut parler plutôt d'archives numériques ou de données numériques. Selon les deux archivistes-paléographes, Françoise Banat-Berger et Christine Nougaret, le terme « donnée » est devenu dominant dans la terminologie législative et réglementaire de divers domaines. En ce qui concerne la terminologie professionnelle archivistique avec ses différentes normes, plusieurs termes ont été utilisés tels que : « record », « document » , « preuve », etc.

[7] Npm. (2023, 11 juillet). *Données numériques - Cabinet NPM| CONSEIL| ETUDE| FORMATION*. Cabinet NPM| CONSEIL| ETUDE| FORMATION. Consulté le 6 avril 2024, à l'adresse https://cabinetnpm.com/donnees-numeriques/

L'avènement de l'ère du numérique, le Big Data et l'intelligence artificielle, a impacté les pratiques de la gestion de l'information et les besoins des utilisateurs.

Ainsi, avec le début d'utilisation des Applications Métiers (AI), des Systèmes d'Informations (SI) et des Systèmes d'Archivage Électronique (SAE), le terme « document » a disparu et ces derniers commencent à générer, stocker et pérenniser des « données » numériques.

> *Le terme de "données" utilisé dans ce cadre est à comprendre dans son acception la plus large, il désigne aussi bien des données non-structurées, semi-structurées ou structurées, brutes ou agrégées, et cela, quels que soient la nature, le métier ou le sujet sur lequel porte ces données...*[8]

En effet, ces données numériques ou archives numériques ne cessent de s'accroître et possèdent la même valeur que des archives papier, tout au long de leur cycle de vie nous devons donc garantir leur intégrité, pérennité, accessibilité, exploitabilité et sécurité.

Ces données peuvent être :

1. Données structurées

Le terme « données structurées » désigne l'ensemble des données quantitatives qui ont été créées sous un modèle bien défini. Dans les bases de données, l'ensemble des données structurées sont stockées. Elles sont généralement constituées par des noms, des dates et des numéros qui sont facilement traités et interprétés par les êtres humains ainsi que les machines.

2. Données non-structurées

Le terme « données non-structurées » désigne l'ensemble des données qualitatives, qui n'ont pas de modèle défini ou structuré. Il s'agit généralement de textes, d'images et de vidéos.

[8] Banat-Berger, F., &Nougaret, C. (2014). Faut-il garder le terme archives ? Des « archives » aux « données ». La Gazette des archives, 233(1), 7–18. Consulté le 9 avril 2024, à l'adresse https://doi.org/10.3406/gazar.2014.5121

3. Données semi-structurées

Le terme « données semi-structurées » désigne l'ensemble des données qui sont facilement interprétées et accessibles par l'être humain ou les machines, mieux que les données non-structurées. En revanche, ces données semi-structurées ne sont pas créées selon des modèles bien définis tels que les données structurées.[9]

III. LA TYPOLOGIE DES DONNEES NUMERIQUES

L'archivage de ces données numériques est une étape très importante puisqu'elles présentent un enjeu crucial pour l'intelligence artificielle qui paraît elle-même nécessaire de nos jours pour assurer le développement dans différents secteurs. En effet, il existe deux types principaux de données numériques à archiver, à savoir :

1. Données chaudes

Les données chaudes, également appelées données dynamiques, sont constituées des données qui sont fréquemment utilisées ou consultées, telles que les documents d'activités au sein d'une entreprise, qui doivent être accessibles pour répondre aux besoins au moment opportun. Ces données sont généralement stockées dans les GED ou les SAE, puisqu'ils facilitent l'accessibilité à ces données au moment opportun et garantissent également leur intégrité et sécurité toute leur durée de vie avec des conditions favorables.

2. Données froides

Contrairement aux données chaudes, les données froides ne sont pas fréquemment utilisées ou consultées. Elles occupent principalement une valeur historique ou de référence, or ceci n'empêche pas que ces données doivent être pérennisées dans des entrepôts ou des centres de données dans des conditions favorables sur des supports de stockage de grande capacité et de faible coût tels que les bandes magnétiques.

---

[9] *Que sont les données structurées ?: Guide complet sur les données structurées*. (s. d.). Elastic. Consulté le 9 avril 2024, à l'adresse https://www.elastic.co/fr/what-is/structured-data

Grâce à un échange téléphonique avec Monsieur Thomas VAN DE WALLE le directeur du numérique et de la conservation aux Archives Nationales de France[10], nous savons que les données froides sont conservées avec un stockage à froid sur des bandes LTO (*Linear Tape-Open*) sur des étagères qui ne nécessitent aucune alimentation électrique.

En revanche, le temps d'accès au contenu de ces bandes est plus long que pour les bibliothèques de sauvegarde ou les librairies de bandes automatisées, bien qu'elles soient plus coûteuses. Cependant, des robots permettent une manipulation automatique et rapide, ce qui facilite ensuite l'accès et la gestion des bandes.[11]

IV. LOCALISATION DES DONNEES NUMERIQUES :

Parallèlement à la multiplication des supports de stockage et à l'accroissement massif des données numériques dans divers domaines à travers le monde, les obligations et les défis des records managers et des archivistes ont aussi évolué. En outre, les données numériques, ou archives numériques, ont des exigences comme les archives papiers pour qu'elles puissent jouer leur rôle de preuve et rester intègres et authentiques.

En effet, les différentes missions ou actions qui donnent naissance à des documents ou des records qui ont une valeur et qui doivent être conservés sur le long terme invitent les records manager ou les archivistes à penser généralement à trois grandes problématiques lors de l'archivage numérique :

- Le format
- Le stockage
- Les normes et les standards

[10] T. Vande Walle (2024, 01 janvier) , communication personnelle. (*cf.* Annexe 6)

[11] Contributeurs aux projets Wikimedia. (2023, 30 avril). *Librairie de sauvegarde*. Consulté le 9 avril 2024, à l'adresse https://fr.wikipedia.org/wiki/Librairie_de_sauvegarde

Selon l'Unesco, la pérennisation du patrimoine numérique est un acte fondamental[12]. Pour ce faire, la mise en place d'une politique d'archivage, de records management ou documentaire, quelle que soit sa dénomination[13], est nécessaire pour obtenir les meilleures conditions de conservation pour toute la durée de vie des documents ou des données numériques. Ceci va permettre de définir : quel document doit-on conserver ? pour combien de temps ? pour quel objectif ? sur quel support et dans quel format ? quel logiciel doit-on choisir ? quel plan de sauvegarde faut-il mettre en place ? quelle procédure de gestion doit-on instaurer ? *etc.*

Cette politique va ensuite permettre de mettre en place des systèmes de veille afin de rester à jour en cas d'évolution technologique de dégradation des supports et des formats, cas dans lesquels il sera nécessaire d'effectuer des opérations de récupération et de migration pour éviter la perte des données numériques.[14]

En effet, selon la politique choisie et les objectifs de pérennisation des données numériques, un choix se fait entre des systèmes de gestion électronique de documents (GED) et des systèmes d'archivage électronique (SAE).

En revanche, dans le domaine de l'archivage électronique à long terme, le choix d'un SAE est la solution la plus fiable, puisqu'elle garantit pendant toute la durée de vie des documents leur intégrité, authenticité, sécurité et exploitabilité. Ces SAE sont généralement hébergés dans des serveurs à forte capacité qui peuvent se trouver au sein de centres de données. En effet, ces serveurs présentent les infrastructures informatiques des SAE et facilitent également la conservation et l'accessibilité des données en toute sécurité.

Au sein de l'Institut Nationale de l'Audiovisuel (INA), nous pouvons trouver les grands serveurs de la France, puisqu'elle gère depuis les années 2000 avec la Bibliothèque nationale de France (BnF), le dépôt légal du web qui collecte environ

---

[12]Ott, F. (2021). *La gestion documentaire au coeur des processus d'affaires : Valider, protéger, exploiter et pérenniser l'information dans l'environnement numérique*. ISTE Group.

[13] Jules, A. (2012). Une politique de gestion des documents d'activité pour une gouvernance documentaire stratégique. *˜la œGazette des Archives/Gazette des Archives*, *228*(4), 153-171. Consulté le 9 avril 2024, à l'adresse www.persee.fr/doc/gazar_0016-5522_2012_num_228_4_4991, https://doi.org/10.3406/gazar.2012.4991

[14]Ott, F. (2021). *La gestion documentaire au coeur des processus d'affaires : Valider, protéger, exploiter et pérenniser l'information dans l'environnement numérique*. *Op. Cit.*

10 To par jour de nouvelles données[15]. Il existe aussi le Centre de Calcul de l'IN2P3 (CC-IN2P3) spécialisé dans la physique nucléaire et la physique des astroparticules, qui contient des milliers de serveurs qui peuvent stocker environ 340 Po de données sur des bandes magnétiques.[16]

V. QUELS SONT LES DEFIS DU STOCKAGE DES DONNEES NUMERIQUES ?

« L'archivage électronique est l'ensemble des actions, outils et méthodes mis en œuvre pour conserver à moyen et à long terme des informations numériques dans le but de les rendre accessibles et exploitables »[17].

Ainsi, l'archivage des données numériques a des exigences similaires à celles des archives papiers. Il faut garantir leur pérennisation à long terme, c'est-à-dire : leur intégrité et fiabilité, leur authenticité, leur exploitabilité et leur interopérabilité. Ces principes des données numériques présentent des défis importants pour les archivistes.

Selon le Centre Informatique National de l'Enseignement Supérieur (CINES), l'obsolescence technologique constitue un *phénomène inéluctable*[18] dans l'archivage des données numériques, puisque que celui-ci repose sur une technologie très évolutive qui rend certains supports de stockage ou formats de données inaccessibles, ce qui pose une grande problématique pour la préservation et la pérennisation des données numériques.

---

[15]INA Institut. (2020, 18 septembre). *A l'ombre des serveurs. Chapitre 3 : capter INA* [Vidéo]. YouTube. Consulté le 10 avril 2024, à l'adresse https://www.youtube.com/watch?v=QI24fNV6ZbM

[16] *Le CC-IN2P3*. (s. d.). Consulté le 10 avril 2024, à l'adresse https://cc.in2p3.fr/qui-sommes-nous/le-cc-in2p3/

[17]*L'archivage électronique [Politique d'archivage]*. (s. d.). Consulté le 22 avril 2024, à l'adresse https://documentation.unistra.fr/Service_Archives/PolitiqueArchivage/co/716_archivageElectronique.html

[18] Dossier « Archivage numérique pérenne », La Gazette du CINES, février 2013 Consulté le 13 avril 2024, à l'adresse https://www.cines.fr/wp-content/uploads/2013/12/Archivage_perenne_Gazette20.pdf

1. L'obsolescence technologique des formats et de l'encodage

Les supports numériques ont, certes, facilité le partage et l'échange des informations entre les systèmes d'information et garantissent l'accès tout au long du cycle de vie de ces informations, mais ceci nécessite des standards essentiels qui définissent :

- Le type d'encodage (ASCII).
- Les formats d'interprétation (PDF, TIFF.).

En effet, ces derniers sont toujours en cours de développement parallèlement au développement technologique, ce qui rend certains formats obsolètes et rend ensuite l'accès aux informations encodées difficile voire impossible.[19]

2. L'obsolescence technologique des supports

En 2023, Bush écrit : "A record if it is to be useful to science, must be continuously extended, it must be stored, and above all it must be consulted."[20] Ainsi, les supports de stockage présentent un élément crucial pour l'exploitation des données numériques. Depuis des années, les supports numériques ont été les plus utilisés pour le stockage grâce à leur facilité d'utilisation et notamment de partage.

En outre, ces derniers ont une durée de vie résistante dans le temps, ainsi que des systèmes de maintenance. Il est donc possible de les rectifier. Les autres supports au contraire, les supports analogiques par exemple, qui sont très fragiles et se dégradent plus rapidement, sont difficiles à réparer et, en cas de dégradation avec le temps, cela peut entraîner l'altération du contenu des supports, voire leur perte.

[19] *Idem.*

[20] Bush, V. (2023, 14 novembre). As we may think. *The Atlantic*. Consulté le 13 avril 2024, à l'adresse https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/

En plus des problèmes d'obsolescence des supports de stockage, nous sommes toujours confrontés aux problèmes d'obsolescence des lecteurs de ces supports, tels que les lecteurs des disquettes.[21]

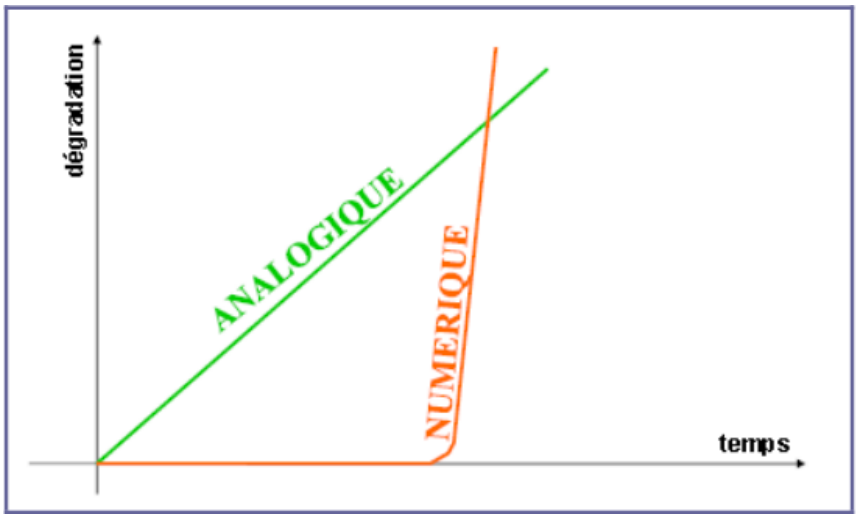

Figure 1 : Comparaison de la dégradation des supports analogiques et les supports numériques en fonction du temps[22] [23]

(Source : L. Duplouy, 2009, cité dans, Hulstaert, 2010)

VI. COMMENT GARANTIR L'INTEGRITE, LA FIABILITE ET LA SECURITE A LONG TERME ?

Afin de garantir l'intégrité, la fiabilité et la sécurité des données numériques à long terme, nous avons besoin de supports de stockage et de formats d'encodage ayant une longue durée de vie. Pour ce faire, il existe des critères pour qu'un support de stockage ou un format soit pérenne.

[21] Dossier « Archivage numérique pérenne », La Gazette du CINES, février 2013 Consulté le 13 avril 2024, à l'adresse https://www.cines.fr/wp-content/uploads/2013/12/Archivage_perenne_Gazette20.pdf

[22]Hulstaert, A. (2010). *Préservation à long terme de l'information numérique* (010/TRIM1/01). SMALS. Consulté le 13 avril 2024, à l'adresse https://www.smalsresearch.be/download/research_reports/deliverable/digital_preservation.pdf

[23] Duplouy. L. (2009). *Séminaire PIN avril 09.* Cité dans Hulstaert, A. (2010). *Préservation à long terme de l'information numérique* (010/TRIM1/01)

1. Les critères de sélection des formats[24]

Selon la politique de la BnF de préservation à long terme des formats de données, il existe de multiples critères à respecter lors de choix de notre support, qui sont : (voir annexe 1)

- CPO-SOC : Critère de pérennité objectif - Communauté d'utilisateurs /Sociabilité
- CPO-DOC : Critère de pérennité objectif - Documentation
- CPO-LIB : Critère de pérennité objectif - Liberté d'utilisation
- CPO-AUT : Critère de pérennité objectif - Indépendance /autonomie
- CPO-ROB : Critère de pérennité objectif - Robustesse
- CPO-COM : Critère de pérennité objectif - Compacité
- CPO-OUT : Critère de pérennité objectif - Disponibilité d'outils de traitement
- CPO-ADD : Critère de pérennité objectif - Contenu Additionnel embarqué
- CPO-PRO : Critère de pérennité objectif - Mécanismes de Protection
- CPO-SIM. : Critère de pérennité objectif - Simplicité
- CPO-STA : Critère de pérennité objectif - Stabilité /Evolutivité
- CPO-TRA : Critère de pérennité objectif - Transparence

Dans l'archivage numérique, l'analyse des formats présente une composante essentielle qui se fait *via* des logiciels différents aidant à montrer si les formats sont plus ou moins utilisés, s'il est plus ou moins facile de les préserver et de les rendre accessibles.

Dans le tableau ci-dessous se trouve la liste de ces outils ainsi que leur utilité.

Nom d'outils	Utilité
Mediainfo / Bitcurator / DPL	Comprendre à quel format appartient un document
TIKA/ DROID/ Seigfried/ FIDO	Identification des formats

[24] Caron, B. (2021). *Formats de données pour la préservation à long terme : la politique de la BnF*. Consulté le 20 avril 2024, à l'adresse https://bnf.hal.science/hal-03374030

VeraPDF / JHOVE / JPLYZER	Validation des formats

Tableau 1 : Les outils d'identification, validation et caractérisation des formats

La grande diversité des données numériques explique la diversité des formats. Il existe un format qui convient pour chaque donnée numérique, tels que les formats généralistes (PDF,ODF, *etc.*) ou les formats métiers comme FITS (Flexible Image Transport System) propres à un domaine défini.[25]

2. Les critères de sélection des supports de stockage

En ce qui concerne les supports de stockage des données numériques, il y a trois critères principaux à prendre en considération, qui sont :

- Le type de support
- La capacité de stockage
- La durée de vie de support
- La sécurité et la protection des données
- L'accès

En outre, il est nécessaire de mettre en place des stratégies pour garantir une qualité optimale de stockage, en créant des copies de sauvegarde par exemple,

> *« La stratégie de stockage de la plateforme de préservation numérique du CINES inclut quatre copies d'un document à préserver, deux sur des baies de disques, deux sur une librairie de bandes. Une de ces quatre copies est externalisée à une distance suffisamment grande pour résister aux catastrophes naturelles. »* (ROUCHON, 2011)

[25] *Chapitre 6 : Les normes et standards utilisés pour l'archivage numérique*. (s. d.). PIAF Portail international archivistique francophone. Consulté le 20 avril 2024, à l'adresse https://www.piaf-archives.org/sites/default/files/bulk_media/m07s04/co/section4_6.html

Réaliser des audits réguliers et mettre en place des procédures de veille, nous permettra de prévenir les risques et de déterminer si nous devons effectuer des migrations aux données conservées ou non[26].

Selon *l'Abrégé de l'archivistique*, dans sa 4ème édition de 2020, nous n'avons pas un support de stockage avec une qualité optimale qui répond parfaitement à nos besoins, car nous faisons toujours face à des défis liés à la dégradation et l'obsolescence. Il est donc important de noter que ces problèmes varient en fonction des types de supports et de leur durée de vie.[27]

Le choix du support se fait en fonction de la période de conservation et du volume des données à conserver. En ce qui concerne les données numériques à durée de conservation limitée, les disques optiques amovibles sont généralement recommandés, mais afin d'éviter la perte de ces données il est conseillé d'effectuer des migrations et de choisir une autre option de stockage en parallèle, c'est-à-dire de créer une deuxième copie afin de garantir la disponibilité des données et d'éviter le risque de perte.

L'inconvénient principal de cette méthode est le coût. En outre, les données numériques de durée de vie moyenne voire longue sont généralement conservées sur des supports fixes tels que les bandes magnétiques et les gros serveurs.[28]

En effet, il existe un ensemble complémentaire des normes et des guides qui fournissent une infrastructure fiable de stockage qui correspondent aux recommandations des meilleures stratégies technologiques de stockage qui permettent de garantir l'interopérabilité et la sécurité des données numériques conservées :

- Les normes IEEE (Institute of Electrical and Electronics Engineers) qui définissent l'architecture des SI.

---

[26] Rouchon, O. (2011, mai). *La démarche qualité au CINES pour la préservation à long-terme des données numériques*. Journées d'Informatique Musicale, Saint-Etienne, France. HAL. Consulté le 20 avril 2024, à l'adresse https://hal.science/hal-03104752/document

[27] *Abrégé d'archivistique : Principes et pratiques du métier d'archiviste*. (2020).

[28] Ott, F. (2021). *La gestion documentaire au coeur des processus d'affaires : Valider, protéger, exploiter et pérenniser l'information dans l'environnement numérique*. *Op.cit.*

- ISO 9660 pour les CD-ROM
- ISO 13962 pour le Digital Linear Tape (DLT)[29]

VII. NOUVELLES TECHNOLOGIES DE STOCKAGE :

Avec l'augmentation des quantités d'archives et les problématiques liées à leur conservation, les chercheurs ont essayé d'inventer de nouvelles technologies de stockage afin de réduire les limites des technologies existantes. Parmi ces nouvelles technologies il existe les plaques de quartz qui ont constitué une révolution du fait de leur durabilité et de leur grande densité.

Cette technologie a été explorée par Microsoft dans le cadre du projet Silica en 2019, qui a permis de stocker environ 7 To de données dans un support de la même taille qu'un DVD et qui peut garantir la disponibilité de ces données sans perte de qualité pour 10 000 ans.

Figure 2 : Comparaison entre les deux supports[30]

En outre, il existe aussi une autre technologie émergente de stockage de données sur de l'acide DésoxyriboNucléique (ADN), le support de l'information

[29] *Chapitre 6 : Les normes et standards utilisés pour l'archivage numérique*. (s. d.). PIAF Portail international archivistique francophone. *Op. Cit.*

[30] Manceau, G. (2023, 18 octobre). *Oubliez le SSD, cette technologie peut stocker des données pendant 10 ; 000 ans ; !* 01net.com. Consulté le 25 avril 2024, à l'adresse https://www.01net.com/actualites/7-to-de-donnees-dans-une-plaque-de-quartz-pour-10-000-ans-le-projet-fou-de-microsoft.html

génétique, qui possède une densité et une durée de vie incomparable par rapport aux autres technologies.

Après avoir expliqué les stratégies liées à l'archivage et la préservation des données numériques à long terme, nous allons explorer dans la deuxième partie du mémoire les principes de base de stockages des données sur les brins d'ADN depuis les premières utilisations jusqu'à nos jours.

PARTIE 2 : LE STOCKAGE SUR ADN

I. QU'EST-CE QU'UN ADN ?

L'acide DésoxyriboNucléique (ADN) est une molécule présente dans toutes les cellules du corps de l'être humain qui sont environ 37 200 milliards de cellules. L'ADN se trouve dans les chromosomes, qui se situent dans le noyau de chaque cellule et comporte toutes les informations relatives à l'activité, l'organisation et la croissance du corps humain.

En outre, l'ADN est créé par environ 20 000 gènes. Il est composé par une double hélice marquée par deux bases azotées : la purine et la pyrimidine.[31]

Ces hélices sont composées de quatre nucléotides complémentaires : l'adénine, la thymine, la guanine et la cytosine. Ils sont connus également sous les dénominations : A, T, C et G.

Les nucléotides sont les composants principaux des acides nucléiques tels que l'ADN. En outre, une composition de 15 à 30 nucléotides donne naissance à des oligonucléotides appelés aussi courts segments d'ADN. [32]

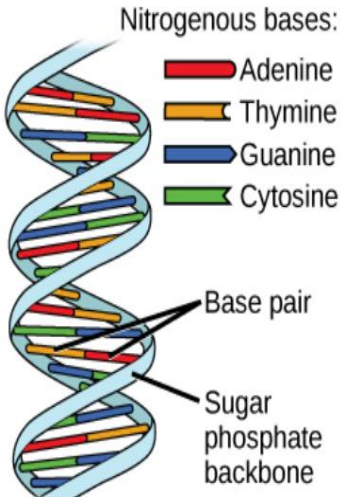

Figure 3: Image explicative de la composition d'ADN

[31] *Les notions-clés de la génétique médicale*. (s. d.). Génétique médicale : ADN, hérédité, tests - Agence biomédecine. Consulté le 28 avril 2024, à l'adresse https://www.genetique-medicale.fr/la-genetique-l-essentiel/les-notions-cles-de-la-genetique/article/les-notions-cles-de-la-genetique-medicale

[32] *Ibid.*,

II. L'ADN : UN SUPPORT DE STOCKAGE DE DONNEES ?

1. Codage de données sur ADN

Avec le développement technologique et la forte utilisation du numérique, les données numériques commencent à s'accroître. Ceci a invité les biologistes à réfléchir à créer un support qui va garantir la pérennisation de ces derniers. De surcroît, ils se sont appuyés sur « une discipline formelle d'ingénierie biologique » (Lemaire, 2023) apparue au début des années 2000 et appelée la Biologie Synthétique, qui permet la conception et la construction de nouveaux systèmes biologiques ou encore la restructuration des systèmes biologiques déjà existants.

En effet, la biologie synthétique a été appliquée dans divers domaines tels que les matériaux (tissu bio fabriqué, bio plastique), l'agro-alimentaire (colorants), les cosmétiques (parfums, arômes), la santé (vaccins, médicaments) et finalement dans le numérique afin d'assurer un stockage numérique sur des brins d'ADN qui a réussi depuis des millions d'années à transférer l'information génétique d'une génération à une autre.

Cette nouvelle technologie du stockage de données sur l'ADN consiste donc à encoder des informations numériques avec un langage binaire (0 et 1) dans le code génétique avec les bases nucléotidiques quaternaires d'ADN :

- Adénine (A)
- Cytosine (C)
- Thymine (T)
- Guanine (G).

En effet, il existe diverses méthodes d'encodage.

Parmi les plus utilisées, il y a :

- 2Bits/base : A=00 ; C=01 ; T=10 ; G=11
- 1Bit/base: A=C=0 ; T=G=1

Cette opération de codage commence par la conversion des données numériques (images, textes ou même vidéos) en code binaire selon une méthode préalablement choisie vers les quatre nucléotides du codage d'ADN.

Ceci permettra ensuite la création chimique de milliers de brins d'ADN artificiels.[33]

2. Décodage de données sur ADN

Pour faire la lecture de ces brins d'ADN, il suffit de les mettre dans des séquenceurs[34] qui vont convertir les brins d'ADN en code binaire. Ce décodage permet d'obtenir finalement les données codées au départ. L'étape de séquençage est très importante pour retrouver les données encodées.[35]

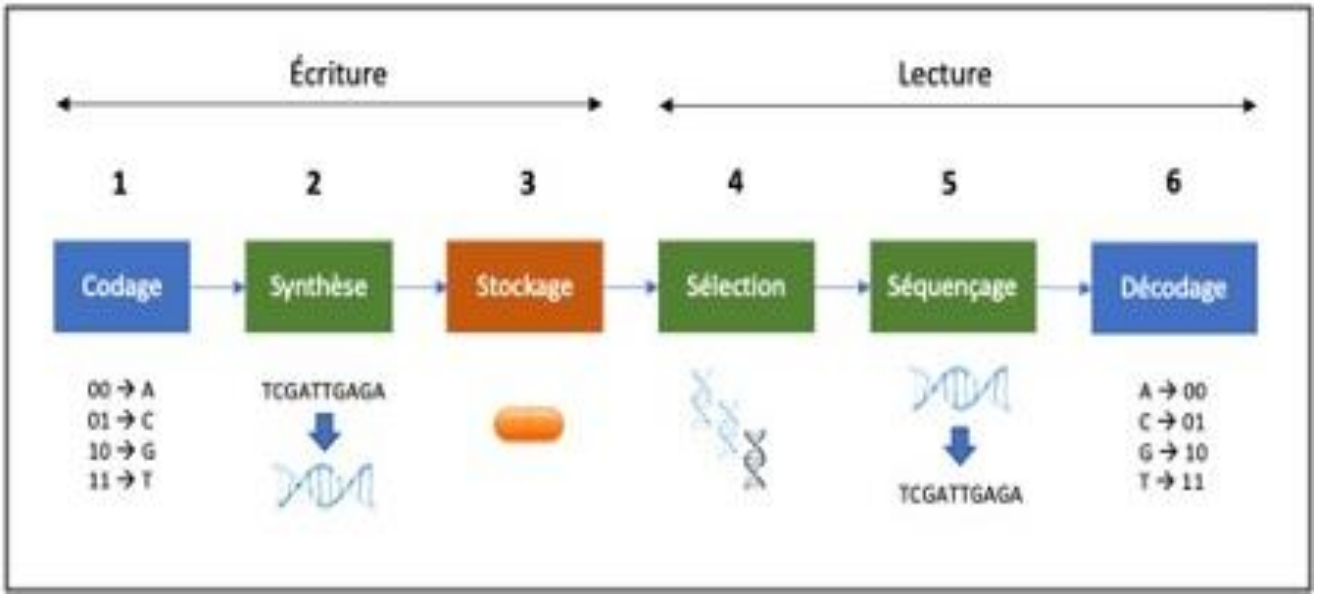

Figure 4 : Codage et décodage des données[36]

---

[33] Sorbonne Université, CNRS, Archives Nationales, & Ministère de la Culture. (2021, 21 novembre). *DNA Drive La Révolution de l'ADN Première mondiale : dépôt d'archives numériques encodées sur ADN aux Archives nationales* [Communiqué de presse]. Consulté le 28 avril 2024, à l'adresse https://www.cnrs.fr/sites/default/files/press_info/202111/CP_R%C3%A9volution_ADN.pdf

[34] Un séquenceur utilisé pour le séquençage d'ADN, c'est-à-dire qu'il permet la lecture et l'analyse des séquences d'ADN.

[35] *Ibid.*

[36] Joanna. (2023, 10 mai). *Stocker les données : la piste prometteuse de l'ADN - Interstices*. Interstices. Consulté le 28 avril 2024, à l'adresse https://interstices.info/stocker-les-donnees-la-piste-prometteuse-de-ladn/

Pour la réussite du processus, nous pouvons distinguer trois générations des séquenceurs :

a. Les séquenceurs de la première génération

Les séquenceurs de la première génération également appelée « Séquenceurs Sanger », apparus en 1986, aident à trouver les nucléotides des petits fragments d'ADN de 500 à 600 paires par base, avec une méthode de séquençage par synthèse de façon automatique avec un taux d'erreur négligeable. En revanche, cette méthode est très coûteuse et le séquençage prend beaucoup de temps.[37]

b. Les séquenceurs de la deuxième génération

Les séquenceurs de la deuxième génération également appelés séquenceurs de nouvelle génération (*Next Generation Sequencing, NGS)*, sont apparus en 2005. Il existe différentes technologies relatives à cette génération tels que Roche 454 (2005), Illumina (2007) et Ion Torrent (2010).

Ces séquenceurs aident à trouver les nucléotides des petits fragments d'ADN (300 bases pour Illumina, 700 bases pour Roche et 400 bases pour Ion Torrent) avec un taux d'erreur de 1% qui se présentent sous forme d'insertion ou de délétion. Contrairement à la première génération, ils ont réussi à réduire le coût et le temps du séquençage.[38]

c. Les séquenceurs de la troisième génération

Les séquenceurs de la troisième génération apparus en 2011, ont réussi à résoudre les problèmes des deux dernières générations. Ils ont augmenté la vitesse du séquençage et ont donc réduit les coûts. En revanche, le taux d'erreur avec cette génération est compris entre 10% et 30%. Il existe différentes technologies

[37] Morisse, P. (2019*). Correction de données de séquençage de troisième génération* [Thèse de doctorat, Normandie Université]. Consulté le 28 avril 2024, à l'adresse https://tel.archives-ouvertes.fr/tel-02320413

[38] *Ibid.*

relatives à cette génération telles que Pacific Biosciences (2011) et Oxford Nanopore Technologies (2014).[39]

III. LES PREMIERES PENSEES DU STOCKAGE SUR SUPPORT D'ADN

➢ L'origine du stockage sur ADN

En décembre 1959, la conférence « *There's Plenty of Room at the Bottom: An Invitation to Enter a New Field of Physics (« Il y a beaucoup de place au fond : une invitation à entrer dans un nouveau domaine de la physique »)* » du physicien Richard Feynman, a eu lieu pendant la réunion annuelle de la société américaine de physique à Caltech. Lors de cette conférence, Feynman a présenté pour la première fois l'idée de la possibilité de la manipulation de structures physiques, chimiques ou biologiques telles que les atomes. Il a fait allusion à la nanotechnologie qui apparaîtra ensuite vers les années 1980.[40,41]

En revanche, dans le volume 43 du journal *Saturday Review* publié le 2 avril 1960, l'article « *The Wonders That Await a Micro-Microscope* » de Richard Feynman aborde l'idée d'un stockage des informations dans des espaces minuscules en s'appuyant sur l'exemple des molécules d'ADN, qui présentent une petite fraction d'une cellule et qui peuvent contenir toutes les informations nécessaires à l'organisation et au développement du corps humain.[42]

> *"All this information—that the eyes are to be brown, that the hair is to be curly, that there is to be an ability to think, that in the embryo the jawbone*

---

[39] *Ibid.*

[40] Contributeurs aux projets Wikimedia. (2023, août 13). *There's Plenty of Room at the Bottom*. Consulté le 28 avril 2024, à l'adresse https://fr.wikipedia.org/wiki/There%27s_Plenty_of_Room_at_the_Bottom

[41] P. Feynman, R. (1959, décembre). *Plenty of room at the bottom*. The American Physical Society, Pasadena. Consulté le 28 avril 2024, à l'adresse https://web.pa.msu.edu/people/yang/RFeynman_plentySpace.pdf

[42] *Saturday Review 1960-04-02 : Vol 43 Iss 14 : Free Download, Borrow, and Streaming : Internet Archive*. (1960, 2 avril). Internet Archive. Consulté le 28 avril 2024, à l'adresse https://archive.org/details/sim_saturday-review_1960-04-02_43_14/page/46/mode/2up

should first develop with a little hole in the side so that later a nerve can grow through it, and thousands of other such instructions—is contained in a very tiny fraction of the cell, in long-chain DNA molecules where one bit of data is crammed inside fifty atoms (inside our pinhead, we have room to give one bit two and a half times that much space).”

IV. LE DEBUT D'APPLICATION DE LA NOUVELLE TECHNOLOGIE DE STOCKAGE

Dans les années 1990, une première démonstration de stockage sur ADN a été développée par « un des pionniers de l'art transgénique »[43], Joe Davis.

De surcroît, il a réussi à créer des *infogènes*[44], pour porter l'intelligence humaine en utilisant la bactérie *Microvenus* [45], présentées via le symbole de superposition des deux lettres Y et I. Cet *infogène* permet de présenter l'information extra biologique avec l'ADN et est susceptible de stocker et de communiquer « des signes tangibles du langage humain de l'ère pré-chrétienne »[46].

En outre, selon Joe Davis, ces *infogènes* se caractérisent par une performance et une robustesse plus forte que les autres machines de communication telles que les ordinateurs par exemple. Le but de cette invention était d'envoyer dans l'espace ces bactéries pour créer des contacts avec des extraterrestres, cela n'a pas été accepté par le comité de biosécurité d'Havard à cause des risques liés à la pollution.

---

[43] Voison, C. (2011). L'art in vivo ou la mythification de la molécule d'ADN. *Images Re-vues*, *8*. Consulté le 28 avril 2024, à l'adresse https://doi.org/10.4000/imagesrevues.503

[44] Un infogène c'est à dire un gène synthétique provenant de la conversion d'une lettre de l'alphabet runique (un alphabet composé de signes magiques et divinatoires) Voison, C. (2020). Art et éthique : L'éthique incertaine des expérimentations de l'art biotechnologique. *Ethica, 24*(1), 17-39. Consulté le 05 mai 2024, à l'adresse https://doi.org/10.3917/jibes.304.0051

[45] Davis, J. (1996). Microvenus. *Art Journal*, *55*(1), 70–74. Consulté le 05 mai 2024, à l'adresse https://doi.org/10.2307/777811

[46] Voison, C. (2020). Art et éthique : L'éthique incertaine des expérimentations de l'art biotechnologique. *Ethica*, *24*(1), 17-39. Consulté le 05 mai 2024, à l'adresse https://doi.org/10.3917/jibes.304.0051

```
01110
00100
00100
00100
00100
00100
```

Figure 5 : Codage du symbole Microvenus

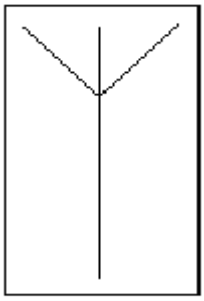

Figure 6: Symbole Microvenus

```
5'-CTTAAAGGGGCCCCCCAACGCGCGCGCT-3'
| | | | | | | | | | | | | | | | | |
5'-CTTAAAGGGGCCCCCCAACGCGCGCGCT-3'
```

Figure 7 : Double brin d'ADN Microvenus formé de 28 nucléotides[47]

En 2012, le chimiste américain George Church a développé avec son équipe, en s'appuyant sur le travail de Joe Davis, une méthode de codage des informations numériques qui lui a permis d'encoder la version HTML du livre *How Synthetic Biology Will Reinvent Nature and Ourselves* de Church GM et Regis E de 659 KB (équivalant à 659 000 octets). Ce livre est constitué de 53 426 mots, 11 images au format JPEG et un programme JavaScript en bit Stream de 5.27 mégabits[48].

Par ailleurs, Church et son équipe ont utilisé les petits fragments d'ADN en simples brins qui sont synthétisés chimiquement, c'est-à-dire qu'ils ne sont pas compatibles avec l'être vivant, ce qui limite ensuite leur duplication. De plus, ces oligonucléotides peuvent occuper environ 200 bases maximum. Ces oligonucléotides ont été amplifiés par PCR (Réaction en Chaine par Polymérase)

---

[47] Voison, C. (2011). *L'art in vivo ou la mythification de la molécule d'ADN*. Images Re-vues, 8. Consulté le 15 mai 2024, à l'adresse https://journals.openedition.org/imagesrevues/503

[48] Church, G. M., Gao, Y., & Kosuri, S. (2012). *Next-Generation Digital Information Storage in DNA*. Science, 337(6102), 1628. Consulté le 15 mai 2024, à l'adresse https://doi.org/10.1126/science.1226355

et ont ensuite été lus par le biais d'un séquenceur « Illumina » afin d'accéder au texte encodé.

De ce fait, l'encodage des bits du livre a donné naissance à environ 54 898 oligonucléotides, c'est-à-dire 159 nucléotides. Or dans un oligonucléotide, ils ont encodé des données de 96 bases dites indexes, qui permettent de donner l'ordre pour la relecture de séquences et de 22 autres bases situées dans les deux extrémités. Church et son équipe ont réussi à montrer la forte densité de stockage de ces petits fragments d'ADN, qui représentent environ 700 téraoctets par millimètre cube.

Ceci constitue le plus grand volume de données stockées sur un support de stockage, à savoir l'ADN[49]. Dans l'article de Church et son équipe, l'image ci-dessous a montré la différence énorme de la densité du support d'ADN et des autres supports utilisés auparavant pour le stockage tels que « l'holographie quantique électronique »[50].

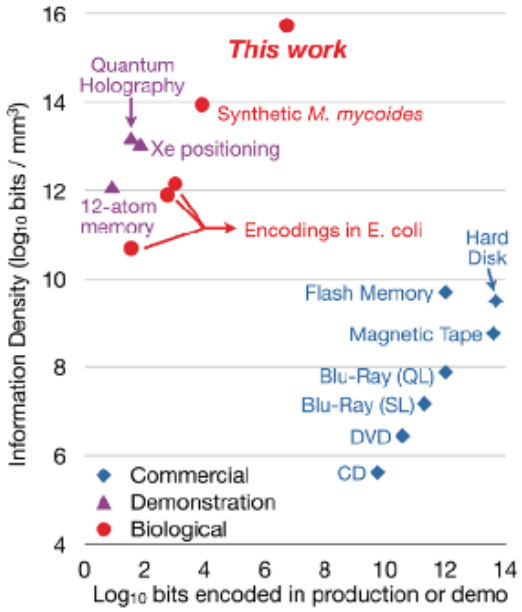

Figure 8 : Image comparative de la densité des supports de stockage[51]

[49] Dimopoulou, M., & Antonini, M. (2022). Data and image storage on synthetic DNA : existing solutions and challenges. EURASIP Journal on Image and Video Processing, 2022(1). Consulté le 15 mai 2024, à l'adresse https://doi.org/10.1186/s13640-022-00600-x

[50] Technologie de stockage d'information d'une densité de 35 bit par électron

[51] Church, G. M., Gao, Y., & Kosuri, S. (2012). Next-Generation Digital Information Storage in DNA. opcit

En 2013, le biologiste Nick Goldman et son équipe de l'Institut Européen de Bio-informatique au Royaume-Uni ont réussi à encoder des informations de différents formats sur des brins d'ADN. Ce travail a permis de transmuter une photographie de leur institut, un texte, des sonnets de Shakespeare et un fichier audio du discours de Martin Luther King « I have a dream », qui représentent au total environ 0.7 Mo. De surcroît, selon Nick Goldman, les résultats obtenus après le décodage de ces fichiers étaient identiques aux fichiers originaux avec un taux d'erreur équivalent à 0%.[52]

Par la suite, plusieurs autres tentatives ont été mises en place afin de prouver la faisabilité de stockage des données sur les brins d'ADN entre 2013 et 2018. Elles ont été présentées par Stéphane Lemaire pendant la conférence « Stockage Numérique : La Révolution de l'ADN ».[53]

[52] SLICE Peuples. (2021, 5 mars). *L'ADN, un média de stockage pour le « Big Data»?SLICE* [Vidéo].YouTube. Consulté le 15 mai 2024, à l'adresse https://www.youtube.com/watch?v=Ev0Z7d3f4UY

[53] Institut d'Astrophysique de Paris. (2022b, mai 24). *« STOCKAGE NUMÉRIQUE : LA RÉVOLUTION DE L'ADN »* [Vidéo]. YouTube. Consulté le 15 mai 2024, à l'adresse https://www.youtube.com/watch?v=BKtKISW_j7o

Total de donnée	Contenu codé	Méthode du synthèse	Longueur en nt	Bits par nucléotide	Année	Travail de
83Ko	Texte de Swiss Federal Charter de 1291 et *The Method of Archimedes*	Phosphoramidite	158	0.86	2015	Dr. Robert N. Grass
151Ko	3 Images en format JPEG	Phosphoramidite	120	0.57	2015	Dr. Robert N. Grass
2Mo	Texte, fichier SVG, PDF, Vidéo, ZipBomb,	Phosphoramidite	152	1.18	2016	Bronholt et Al.
22Mo	Film MPEG	Phosphoramidite	230	0.89	2016	Bronholt et Al.
200.2Mo	35 fichiers compris : vidéo, images, audio, texte (*la Délaration universelle de droits de l'Homme*) traduit en 100 langues	Phosphoramidite	150-200	0.81	2018	Organick et Al
8.5Mo	Fichier texte ; *Bible* en anglais et en Hébreu	Phosphoramidite	194	1.94	2018	Anavy et Al.
854 o	Fichier texte	Phosphoramidite	85	1.78	2018	Choi et Al.

Tableau 2 : Tableau récapitulatif des tentatives de stockage sur ADN[54]

[54] *Idem.*

PARTIE 3 : ÉTUDE DE FAISABILITE DU STOCKAGE SUR DES BRINS D'ADN : CAS DE L'ADN BIOCOMPATIBLE DE BIOMEMORY

I. PRESENTATION DE BIOMEMORY

BIOMEMORY est une startup fondée en juillet 2021 à Paris suite à la collaboration entre le directeur de recherche au Centre National de la Recherche Scientifique (CNRS) Stéphane Lemaire, le maître de conférences à la Sorbonne Université Pierre Crozet et l'entrepreneur des technologies avancées Erfane Arwani[55]. La nouvelle technologie révolutionnaire de stockage des données numériques sur des molécules d'ADN biocompatibles a été inventée au sein de BIOMEMORY. L'idée de lancement de ce projet a vu le jour en 2018, suite à un défi avec les étudiants de l'association Alma mater pour encoder la *Déclaration des droits de l'homme et du citoyen*. Stéphane Lemaire et Pierre Crozet ont réussi à l'encoder et ont également encodé la *Déclaration de droits de la femme et de la citoyenne* sur deux capsules d'ADN conservées dans l'armoire de fer des Archives nationales.

Figure 9 : Les capsules des deux Déclarations[56]

[55] Tech, S. (2024, 5 avril). *Nous voulons rendre l'informatique moléculaire accessible et pratique.* Consulté le 17 mai 2024, à l'adresse https://fr.linkedin.com/pulse/nous-voulons-rendre-linformatique-mol%C3%A9culaire_ga5fe?trk=public_post_reshare_feed-article-content

[56] *Le stockage des données sur l'ADN : une technologie révolutionnaire*. (2022,30 juin). Sorbonne Université. Consulté le 17 mai 2024, à l'adresse https://sciences.sorbonne-universite.fr/actualites/le-stockage-des-donnees-sur-ladn-une-technologie-revolutionnaire

II. EXPLICATION DE LA TECHNOLOGIE DE STOCKAGE SUR L'ADN CHEZ BIOMEMORY

L'ADN constitue le support de stockage de l'information le plus stable dans le monde. Grâce à l'analyse de l'ADN présent dans des fossiles de mammouth laineux trouvés dans le pergélisol sibérien, les scientifiques ont pu découvrir plusieurs informations sur ces animaux telles que leur régime alimentaire et leur apparence.[57]

En effet, durant mon entretien avec le chercheur Pierre Crozet, celui-ci m'a expliqué cette nouvelle technologie de stockage des données numériques sur de l'ADN. Tout d'abord, cette technologie consiste à stocker des données informatiques en 0 et 1, qui sont traduits en format d'ADN en ACTG via une façon de codage. Au sein de BIOMEMORY, ils utilisent la technique de George Church de 1Bit par base, c'est-à-dire, le 0 est traduit par A ou C et le 1 est traduit en T ou G par l'ordinateur.

Ensuite, une fois qu'ils ont créé la séquence, ils la synthétisent chimiquement via une méthode qui s'appelle la chimie Phosphoramidite[58]. Après, ils obtiennent une molécule synthétisée sous format physique qui peut être stockée sur deux types de supports après leur déshydratation. Ces supports, qui n'ont pas la même vocation, sont les suivants :

- Des capsules développées par Imagène.
- Des DNA-Card développées par BIOMEMORY.

L'objectif de ces supports est de protéger l'ADN contre trois facteurs essentiels qui sont l'eau, l'oxygène et la lumière. Après la mise en place des molécules synthétisées dans les capsules, l'équipe d'Imagène ajoute de l'argon et

[57] Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., & Schuster, S. C. (2006). Metagenomics to Paleogenomics : Large-Scale Sequencing of Mammoth DNA. *Science*, *311*(5759), 392-394. Consulté le 22 mai 2024, à l'adresse https://doi.org/10.1126/science.1123360

[58] Il s'agit c'est une méthode de chimie qui utilise des bases qui sont produites par chimie de synthèse, en gros, par du pétrole

les scelle avec le laser afin de les fermer complètement, ceci garantissant la stabilité de l'ADN à l'intérieur pour 50 000 ans.

En outre, pour accéder au contenu de l'ADN, il suffit de choisir le séquenceur correspondant au support (capsules, Card, etc.) utilisé pour le stockage afin de récupérer et de lire l'ADN.

III. AVANTAGES ET LIMITES DE STOCKAGE SUR LES SUPPORTS D'ADN

Les défis liés à la préservation des données numériques ne cessent pas de croître et qui devraient atteindre environ 175Zo en 2025, de plus, les désavantages des supports de stockage actuels ont une durée de vie très limitée. En outre, les disques SSD, les bandes magnétiques et les bandes LTO doivent être changés tous les 5 ou 7 ans en moyen, cependant, ces supports sont accumulés dans les centres des data qui occupent une énorme surface dans le monde, environ 167Km^2.

De plus, la conservation de ces supports nécessite des conditions bien définies afin de les préserver, mais cela contribue fortement à l'augmentation de l'empreinte carbone qui dépasse celle de l'aviation. Ainsi, leur consommation d'électricité vaut environ 2% de la consommation mondiale. Les supports de stockage actuels sont donc des supports énergivores, volumineux et fragiles.[59]

Pour résoudre ce problème, le support de stockage d'ADN présente une alternative fiable et stable pour le stockage des données numériques pour une centaine de milliers d'années, ce qui répond à la problématique d'obsolescence des supports. Ensuite, les supports d'ADN proposent une grande capacité de stockage, leur densité maximale étant 4,5*10^{20} octets, ce qui veut dire que toutes

[59] Ministère de la culture, Sorbonne Université, & Archives Nationales. (2021, 20 novembre). *DNA Drive la révolution de l'ADN première mondiale : dépôt d'archives numériques encodées sur ADN aux archives nationales* [Communiqué de presse]. Consulté le 25 mai 2024, à l'adresse https://www.cnrs.fr/sites/default/files/press_info/2021-11/DP_DNA_Drive_R%C3%A9volution_ADN.pdf

les données produites dans le monde en 2019 pourraient être stockées en 100g d'ADN, « soit le volume d'une tablette au chocolat »[60] (Lemaire, 2022).

De plus, cette technologie de stockage sur ADN ne nécessite pas d'intervention coûteuse en énergie puisqu'il s'agit de molécules stables qui peuvent être conservées à température ambiante.[61]

En revanche, ces supports d'ADN ont deux inconvénients principaux qui sont les coûts et temps d'exploitation, ce qui limite leur utilisation en tant que solution aujourd'hui dans les différents centres de documentation. De plus, un autre inconvénient est l'absence d'un cadre normatif et réglementaire qui fixe l'ensemble des exigences relatives à ce support afin de renforcer la transparence des supports d'ADN en tant que support fiable pour un stockage à long terme des données numériques.

Pour conclure, le SWOT ci-dessous donne une vision d'ensemble des forces, faiblesses, opportunités et menaces liées à ce projet :

Forces	• Durabilité • Stabilité • Non-énergivore • Forte densité
Faiblesses	• Coût • Temps d'exploitation • Complexité technique
Opportunités	• Diminution de l'empreinte du carbone • Gain d'espace
	• Absence d'un cadre réglementaire

[60] Institut d'Astrophysique de Paris. (2022, 24 mai). *« STOCKAGE NUMÉRIQUE : LA RÉVOLUTION DE L'ADN »* [Vidéo]. YouTube. Consulté le 25 mai 2024, à l'adresse https://www.youtube.com/watch?v=BKtKISW_j7o

[61] Maes, A., Peillet, J. L., Julienne, A., Blachon, C., Cornille, N., Gibier, M., Arwani, E., Xu, Z., Crozet, P., & Lemaire, S. D. (2022). La révolution de l'ADN : biocompatible and biosafe DNA data storage. *bioRxiv (Cold Spring Harbor Laboratory)*. Consulté le 30 mai 2024, à l'adresse https://doi.org/10.1101/2022.08.25.505104

Menaces	• Sécurité

IV. FIABILITE, INTEGRITE ET SECURITE DES DONNEES STOCKEES SUR ADN BIOCOMPATIBLE :

Les supports de stockage doivent répondre à certaines exigences afin de garantir dans le temps l'intégrité, l'exploitabilité et la sécurité des données numériques. En revanche, la lecture de l'ADN est destructrice, c'est-à-dire qu'une fois que nous avons accédé au contenu des bras dans une capsule, nous ne pouvons pas les récupérer.

Selon le directeur des Archives Nationales, monsieur Thomas VAN DE WALLE, cette inaccessibilité pose une grande limite pour ce support. En revanche, Pierre Crozet trouve que ce n'est pas une problématique puisque chez BIOMEMORY, il travaille à l'échelle microscopique, c'est-à-dire en utilisant le nombre d'Avogadro[62] égal à $6.022*10^{23}$, ce qui explique le nombre immense des copies de chaque donnée. Donc même s'ils perdent un million de molécules, il reste encore une quantité largement suffisante.

En outre, la technologie utilisée chez BIOMEMORY est celle du stockage sur ADN double brin, aussi nommé DNA Drive. Ils utilisent des molécules d'ADN plus longues, en double brin et compatibles avec la croissance *in vitro et in vivo*[63] contenant deux index aux extrémités qui permettent l'accès aléatoire aux données ainsi que l'assemblage des séquences après le séquençage[64].

Ces molécules sont stockées après un contrôle qualité de chaque synthèse afin d'éviter le risque de mutation et pour être sûr que les informations codées sont intègres. Afin de faciliter la réplication de ces molécules, il existe des

[62] Le nombre d'Avogadro présente le nombre des particules qui se trouve dans une mole. (Mole = quantité de matière d'un système)

[63] Maes, A., Peillet, J. L., Julienne, A., Blachon, C., Cornille, N., Gibier, M., Arwani, E., Xu, Z., Crozet, P., & Lemaire, S. D. (2022). La révolution de l'ADN : biocompatible and biosafe DNA data storage. *bioRxiv (Cold Spring Harbor Laboratory)*. Consulté le 30 mai 2024, à l'adresse https://doi.org/10.1101/2022.08.25.505104

[64] *Ibid*,

méthodologies de réplication, qui peuvent être via la PCR ou via des bactéries telles que la Chériechacolie. Il suffit d'ajouter des signaux pour que la bactérie comprenne ce qu'elle doit copier, il suffit d'ajouter un peu de sucre dessus et au bout de vingt minutes elle va se multiplier[65].

Cependant, selon Pierre Crozet l'intégrité des copies de données numériques codées sur les brins d'ADN comporte un taux faible d'erreur et c'est là où la biologie est particulièrement forte, car par rapport aux autres méthodes de copie, manuellement ou via le matériel informatique, il est facile de faire des erreurs. Or la réplication via les bactéries peut engendrer une erreur pour chaque milliard de base et par la PCR une erreur pour chaque million de base, ce qui reste incomparable avec les autres supports.

En ce qui concerne la sécurité des données et les problèmes de vol par exemple, Pierre Crozet a confirmé qu'il est impossible d'accéder aux contenus sur les brins d'ADN : il s'agit d'une échelle microscopique, aussi sont-ils encodés selon une méthode précise, avec un encodage nécessitant des logiciels spécifiques, qui sont des logiciels propriétaires appartenant à BIOMEMORY. De plus, des techniques de cryptage sont en cours de développement.

Ainsi, ces supports d'ADN ont une durée de vie longue, ce qui élimine la problématique d'obsolescence telle que présente pour les autres supports puisqu'on parle d'une durée de vie de 50 000 ans pour les capsules et de 150 ans pour les DNA-Card, ce qui paraît incomparable aux bandes LTO, par exemple, que nous devons changer tous les 7 ans au moins.

V. QUELLES DONNEES A STOCKER SUR CES NOUVEAUX SUPPORTS ?

Cette nouvelle technologie permet de stocker tout type de données : image, vidéo, texte, *etc*. BIOMEMORY présente le vendeur de ses supports comme visant de futurs clients occupant les gros Data Center qui font partie de leur business, notamment les GAFAM : Google, Apple, Facebook, Amazon, et

[65] Chut. (2022, août 31). *Stockage ADN : la révolution miniature*. *Chut !* Consulté le 30 mai 2024, à l'adresse https://chut.media/tech/stockage-adn-toutes-les-donnees-du-monde-dans-une-tablette-de-chocolat/

Microsoft. En effet, ils fournissent une infrastructure fiable pour le stockage d'une telle masse de données qui ne cesse de croître.

Le Centre National de l'audiovisuel (INA), chargé de la conservation des archives audiovisuelles en France, déclare selon leur rapport d'activité de 2023, 27 498 595 heures de documents radiophoniques et télévisuels conservées, 17 127 sites web audiovisuels et médias et 147 milliards de versions d'URL conservées depuis 1996[66]. Ainsi, la gestion et la conservation de cette quantité des données sur des serveurs peut provoquer plusieurs problèmes techniques liés aux supports de ces données ainsi qu'un fort impact environnemental.

Pour y faire face, l'INA pourrait faire partie des futurs clients de BIOMEMORY. Dès que la technologie d'ADN se démocratisera et deviendra accessible et exploitable par plusieurs centres d'archives ou de documentation, ceci rendra la mise en place d'un cadre réglementaire ou normatif important qui identifie les exigences de ce support.

Et, par la suite, certaines faiblesses de cette technologie sont susceptibles d'être résolues. Cela permettra de renforcer la confiance de son utilisation, de la rendre plus accessible et mieux acceptée par le grand public et, par la suite, de considérer d'un point de vue archivistique ce support comme un support fiable pour une conservation à long terme des données numériques.

VI. ROLE DES ARCHIVISTES

En outre, cette évolution technologique et ce bouleversement touchent essentiellement le domaine des sciences de l'information et le rôle des archivistes qui travaillent afin de garantir dans le temps l'accessibilité et la pérennisation des données numériques pour créer la mémoire. Donc, le rôle des archivistes reste crucial, car c'est l'archiviste qui va prendre la décision, préciser quelles données sont à conserver, pour quelle durée, sur quel support, selon quelles priorités, et veiller au respect du cadre légal et règlementaire des données lors de l'archivage.

[66] INA. (2023). INA : RAPPORT D'ACTIVITÉS 2023. Dans *INA*. Consulté le 5 juin 2024, à l'adresse https://www.ina.fr/hub-p/public/2024-05/EXE_INA%20RA%202023_BAT%20BD3-pdf.pdf

VII. AVENIR DE CETTE TECHNOLOGIE

Les avancements de cette technologie au sein de BIOMEMORY sont en cours de développement. Selon Pierre Crozet, ils ont réussi à créer la nouvelle ère de « l'information moléculaire » et sont désormais en train de se développer dans l'objectif de résoudre les problématiques liées aux temps d'accès et au coût : parmi leurs objectifs fixés pour 2030, on compte le fait de faire baisser le téraoctet à un dollar chez BIOMEMORY.

En somme, le stockage sur ADN ouvre des perspectives faramineuses selon Pierre Crozet, puisqu'il sera possible à terme de transporter les données du monde vers Mars. En effet, dans environ 30 ans, BIOMEMORY devrait terminer ses développements en cours dans le *biocomputing*, c'est-à-dire le fait d'utiliser les systèmes biologiques à la place de l'informatique et de réussir avec des molécules et des bactéries à résoudre des problèmes informatiques. Par la suite, ces bactéries devraient réussir à encoder des données numériques d'une façon autonome *via* des opérations logiques et être ensuite capables de comprendre et de traiter les informations. Ainsi, l'objectif de BIOMEMORY est d'apprendre aux bactéries comment faire les choses et par conséquent, augmenter la vitesse de codage, de lecture et de partage.

CONCLUSION

Au cours de ce mémoire nous nous sommes intéressés au stockage des données numériques, et plus particulièrement à la nouvelle technologie de stockage sur des supports d'ADN. Notre objectif a été de proposer une étude de faisabilité de cette technologie et de déterminer s'il s'agit de supports fiables. Nous souhaitions également comprendre l'utilité de ce support qui peut être utilisé comme un remplaçant dans les différents centres d'archivage afin de limiter les inconvénients techniques et environnementaux liés aux autres supports utilisés aujourd'hui pour le stockage des données numériques.

Afin de clarifier cette problématique, nous avons rappelé dans la première partie l'histoire des différentes générations des supports d'archivage de données numériques. Aussi, nous avons exploré les différentes exigences de choix des formats et des supports garantissant la préservation des données numériques à long terme. Cette première partie a permis d'expliquer la notion de la préservation des données numériques ainsi que les différentes normes à prendre en considération lors de la sélection d'un format ou d'un support afin de faciliter ensuite la migration des données pour éviter leur perte en cas de dommage ou d'obsolescence.

Dans la deuxième partie, nous avons exploré les supports d'ADN, en commençant par l'explication de l'évolution de ces supports d'un point de vue scientifique. Nous avons également expliqué et analysé les différentes étapes à suivre afin de passer d'un langage binaire interprétable par une machine à un code génétique reposant sur les bases nucléotidiques quaternaire d'ADN (ACTG). Ce résumé de l'histoire de l'utilisation de la technologie d'ADN a été suivi d'une présentation des tentatives de différents domaines, en vue de mettre en évidence l'utilité de cette technologie pour le futur stockage.

Enfin, dans la dernière partie de ce mémoire, nous nous sommes concentrés sur la tentative de BIOMEMORY de stockage des données numériques sur des brins d'ADN compatibles, ce qui nous a permis de naviguer dans cette invention et de comprendre ses avantages ainsi que ses inconvénients pour la préservation des données numériques.

Ce mémoire nous a permis de conclure que le support d'ADN constitue une opportunité très intéressante grâce aux divers avantages qui ont contribué à le rendre fiable pour un stockage à long terme, à sa stabilité ainsi qu'à sa forte densité de stockage. Ce support peut être une solution pour des données qui doivent être préservées dans les meilleures conditions, mais qui ne sont plus consultables et qui occupent beaucoup d'espace dans les locaux de conservation. Ainsi, en tant qu'archivistes, l'un de nos objectifs principaux est de rendre l'information pertinente disponible dans des délais très limités afin de garantir la satisfaction des usagers.

En revanche, les inconvénients du support d'ADN sont importants par rapport aux supports utilisés aujourd'hui, et ce support ne peut malheureusement pas remplacer les technologies actuelles telles que les disques optiques et les bandes magnétiques, qui ont une vitesse d'accès et d'écriture des données plus rapide. Ainsi, il existe d'autres nouvelles technologies qui présentent des alternatives fiables avec un coût et un temps d'accessibilité moins élevés que ceux de l'ADN.

Pour conclure, l'ADN est un support du futur. Actuellement, nous sommes malheureusement dans une phase intermédiaire entre les archives physiques et numériques. Or, cette transition a déjà causé des ambiguïtés dans la gestion et la conservation des archives. En outre, cela explique l'une des différentes difficultés majeures que nous avons rencontrées lors de la préparation de ce mémoire, ce qui a limité notre périmètre de recherche. En effet, quelques entretiens avaient été prévus avec différentes personnes, mais ils n'ont pas été menés. Cela a restreint les informations recueillies, car plusieurs entretiens avec différents échantillons provenant de divers domaines auraient pu nous aider à élargir notre vision et à diversifier les différentes conséquences liées à cette technologie. En outre, le stockage sur les supports d'ADN reste un sujet nouveau, ce qui explique le manque d'écrits dans la littérature scientifique dans le domaine des archives numériques.

.

SOURCES

• DOCUMENTS CONSULTES

1. Flermond, F. R. (2017). *Histoire des supports de stockage : de la carte perforée à la clé USB* (ENSSIB, Éd.) [Mémoires Master « Archives numériques », ENSSIB]. . Consulté le 6 avril 2024, à l'adresse https://www.enssib.fr/bibliotheque-numerique/documents/67744-histoire-des-supports-de-stockage-de-la-carte-perforee-a-la-cle-usb.pdf
2. Institut d'Astrophysique de Paris. (2022b, mai 24). *« STOCKAGE NUMÉRIQUE : LA RÉVOLUTION DE L'ADN »* [Vidéo]. YouTube. Consulté le 15 mai 2024, à l'adresse https://www.youtube.com/watch?v=BKtKISW_j7o
3. Morisse, P. (2019*). Correction de données de séquençage de troisième génération* [Thèse de doctorat, Normandie Université]. Consulté le 28 avril 2024, à l'adresse https://tel.archives-ouvertes.fr/tel-02320413
4. Sorbonne Université, CNRS, Archives Nationales, & Ministère de la Culture. (2021, 21 novembre). *DNA Drive La Révolution de l'ADN Première mondiale : dépôt d'archives numériques encodées sur ADN aux Archives nationales* [Communiqué de presse]. Consulté le 28 avril 2024, à l'adresse https://www.cnrs.fr/sites/default/files/press_info/202111/CP_R%C3%A9volution_ADN.pdf

• ENTRETIENS

1. Entretien de BIOMEMORY avec Pierre Crozet
2. Entretien des Archives Nationales avec Van De Walle Thomas
3. Entretien de CC-IN2P3 avec Jean-Yves Nief

BIBLIOGRAPHIE

• ARCHIVAGE NUMERIQUE

1. Banat-Berger, F., &Nougaret, C. (2014). Faut-il garder le terme archives ? Des « archives » aux « données ». La Gazette des archives, 233(1), 7–18. Consulté le 9 avril 2024, à l'adresse https://doi.org/10.3406/gazar.2014.5121
2. Ott, F. (2021). *La gestion documentaire au coeur des processus d'affaires : Valider, protéger, exploiter et pérenniser l'information dans l'environnement numérique*. ISTE Group.
3. Npm. (2023, 11 juillet). *Données numériques - Cabinet NPM| CONSEIL| ETUDE| FORMATION*. Cabinet NPM| CONSEIL| ETUDE| FORMATION. Consulté le 6 avril 2024, à l'adresse https://cabinetnpm.com/donnees-numeriques/
4. *Que sont les données structurées ?: Guide complet sur les données structurées*. (s. d.). Elastic. Consulté le 9 avril 2024, à l'adresse https://www.elastic.co/fr/what-is/structured-data
5. Navarro, R. (2023, 13 octobre). *Donnée « chaude » et donnée « froide » dans le cadre de jumeau numérique*. HEXABIM. Consulté le 9 avril 2024, à l'adresse https://www.hexabim.com/publications/donnee-chaude-et-donnee-froide-dans-le-cadre-de-jumeau-numerique
6. Contributeurs aux projets Wikimedia. (2023, 30 avril). *Librairie de sauvegarde*. Consulté le 9 avril 2024, à l'adresse https://fr.wikipedia.org/wiki/Librairie_de_sauvegarde
7. Jules, A. (2012). Une politique de gestion des documents d'activité pour une gouvernance documentaire stratégique. *˜la Gazette des Archives/Gazette des Archives*, *228*(4), 153-171. Consulté le 9 avril 2024, à l'adresse www.persee.fr/doc/gazar_0016-5522_2012_num_228_4_4991, https://doi.org/10.3406/gazar.2012.4991
8. INA Institut. (2020, 18 septembre). *A l'ombre des serveurs. Chapitre 3 : capter INA* [Vidéo]. YouTube. Consulté le 10 avril 2024, à l'adresse https://www.youtube.com/watch?v=QI24fNV6ZbM
9. *Le CC-IN2P3*. (s. d.). Consulté le 10 avril 2024, à l'adresse https://cc.in2p3.fr/qui-sommes-nous/le-cc-in2p3/

10. *L'archivage électronique [Politique d'archivage]*. (s. d.). Consulté le 22 avril 2024, à l'adresse https://documentation.unistra.fr/Service_Archives/PolitiqueArchivage/co/716_archivageElectronique.html
11. Dossier « Archivage numérique pérenne », La Gazette du CINES, février 2013 Consulté le 13 avril 2024, à l'adresse https://www.cines.fr/wp-content/uploads/2013/12/Archivage_perenne_Gazette20.pdf .
12. Bush, V. (2023, 14 novembre). As we may think. *The Atlantic*. Consulté le 13 avril 2024, à l'adresse https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
13. Hulstaert, A. (2010). *Préservation à long terme de l'information numérique* (010/TRIM1/01). SMALS. Consulté le 13 avril 2024, à l'adresse https://www.smalsresearch.be/download/research_reports/deliverable/digital_preservation.pdf
14. Duplouy. L. (2009). *Séminaire PIN avril 09.* Cité dans Hulstaert, A. (2010). *Préservation à long terme de l'information numérique* (010/TRIM1/01)
15. Caron, B. (2021). *Formats de données pour la préservation à long terme : la politique de la BnF*. Consulté le 20 avril 2024, à l'adresse https://bnf.hal.science/hal-03374030
16. *Chapitre 6 : Les normes et standards utilisés pour l'archivage numérique*. (s. d.). PIAF Portail international archivistique francophone. Consulté le 20 avril 2024, à l'adresse https://www.piaf-archives.org/sites/default/files/bulk_media/m07s04/co/section4_6.html
17. Rouchon, O. (2011, mai). *La démarche qualité au CINES pour la préservation à long-terme des données numériques*. Journées d'Informatique Musicale, Saint-Etienne, France. HAL. Consulté le 20 avril 2024, à l'adresse https://hal.science/hal-03104752/document
18. *Abrégé d'archivistique : Principes et pratiques du métier d'archiviste*. (2020).
19. INA. (2023). INA : RAPPORT D'ACTIVITÉS 2023. Dans *INA*. Consulté le 5 juin 2024, à l'adresse https://www.ina.fr/hub-p/public/2024-05/EXE_INA%20RA%202023_BAT%20BD3-pdf.pdf

- **STOCKAGE SUR ADN**

1. SLICE Peuples. (2021, mars 5). *L'ADN, un média de stockage pour le « Big Data » ? SLICE* [Vidéo]. YouTube. Consulté le 2 avril 2024, à l'adresse https://www.youtube.com/watch?v=Ev0Z7d3f4UY
2. Goldman, Nick & Bertone, Paul & Chen, Siyuan & Dessimoz, Christophe & Leproust, Emily & Sipos, Botond & Birney, Ewan. (2013). Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. Nature. 494. Consulté le 2 avril 2024, à l'adresse https://doi.org/10.1038/nature11875
3. I'MTech. (2022, 18 mars). *L&rsquo ; ADN pour stocker les données*. I'MTech. Consulté le 2 avril 2024, à l'adresse https://imtech.imt.fr/2020/03/25/ladn-pour-stocker-les-donnees-oligoarchive/
4. Karayan, R. (2021, 8 décembre). Les Archives Nationales inaugurent le stockage numérique sur ADN. *www.usine-digitale.fr*. Consulté le 2 avril 2024, à l'adresse https://www.usine-digitale.fr/article/les-archives-nationales-inaugurent-le-stockage-numerique-sur-adn.N1162567
5. Manceau, G. (2023, 18 octobre). *Oubliez le SSD, cette technologie peut stocker des données pendant 10 ; 000 ans ; !* 01net.com. Consulté le 25 avril 2024, à l'adresse https://www.01net.com/actualites/7-to-de-donnees-dans-une-plaque-de-quartz-pour-10-000-ans-le-projet-fou-de-microsoft.html
6. *Les notions-clés de la génétique médicale*. (s. d.). Génétique médicale : ADN, hérédité, tests - Agence biomédecine. Consulté le 28 avril 2024, à l'adresse https://www.genetique-medicale.fr/la-genetique-l-essentiel/les-notions-cles-de-la-genetique/article/les-notions-cles-de-la-genetique-medicale
7. Sorbonne Université, CNRS, Archives Nationales, & Ministère de la Culture. (2021, 21 novembre). *DNA Drive La Révolution de l'ADN Première mondiale : dépôt d'archives numériques encodées sur ADN aux Archives nationales* [Communiqué de presse]. Consulté le 28 avril 2024, à l'adresse

https://www.cnrs.fr/sites/default/files/press_info/202111/CP_R%C3%A9volution_ADN.pdf

8. Joanna. (2023, 10 mai). *Stocker les données : la piste prometteuse de l'ADN - Interstices*. Interstices. Consulté le 28 avril 2024, à l'adresse https://interstices.info/stocker-les-donnees-la-piste-prometteuse-de-ladn/
9. Contributeurs aux projets Wikimedia. (2023, août 13). *There's Plenty of Room at the Bottom*. Consulté le 28 avril 2024, à l'adresse https://fr.wikipedia.org/wiki/There%27s_Plenty_of_Room_at_the_Bottom
10. P. Feynman, R. (1959, décembre). *Plenty of room at the bottom*. The American Physical Society, Pasadena. Consulté le 28 avril 2024, à l'adresse https://web.pa.msu.edu/people/yang/RFeynman_plentySpace.pdf
11. *Saturday Review 1960-04-02 : Vol 43 Iss 14 : Free Download, Borrow, and Streaming : Internet Archive*. (1960, 2 avril). Internet Archive. Consulté le 28 avril 2024, à l'adresse https://archive.org/details/sim_saturday-review_1960-04-02_43_14/page/46/mode/2up
12. Voison, C. (2011). L'art in vivo ou la mythification de la molécule d'ADN. *Images Re-vues, 8*. Consulté le 28 avril 2024, à l'adresse https://doi.org/10.4000/imagesrevues.503
13. Voison, C. (2020). Art et éthique : L'éthique incertaine des expérimentations de l'art biotechnologique. *Ethica, 24*(1), 17-39. Consulté le 05 mai 2024, à l'adresse https://doi.org/10.3917/jibes.304.0051
14. Davis, J. (1996). Microvenus. *Art Journal*, *55*(1), 70–74. Consulté le 05 mai 2024, à l'adresse https://doi.org/10.2307/777811
15. Church, G. M., Gao, Y., & Kosuri, S. (2012). *Next-Generation Digital Information Storage in DNA*. Science, 337(6102), 1628. Consulté le 15 mai 2024, à l'adresse https://doi.org/10.1126/science.1226355
16. Dimopoulou, M., & Antonini, M. (2022). Data and image storage on synthetic DNA : existing solutions and challenges. EURASIP Journal on

Image and Video Processing, 2022(1). Consulté le 15 mai 2024, à l'adresse https://doi.org/10.1186/s13640-022-00600-x

17. SLICE Peuples. (2021, 5 mars). *L'ADN, un média de stockage pour le « Big Data»?SLICE* [Vidéo].YouTube. Consulté le 15 mai 2024, à l'adresse https://www.youtube.com/watch?v=Ev0Z7d3f4UY
18. Tech, S. (2024, 5 avril). *Nous voulons rendre l'informatique moléculaire accessible et pratique.* Consulté le 17 mai 2024, à l'adresse https://fr.linkedin.com/pulse/nous-voulons-rendre-linformatique-mol%C3%A9culaire ga5fe?trk=public_post_reshare_feed-article-content
19. *Le stockage des données sur l'ADN : une technologie révolutionnaire*. (2022,30 juin). Sorbonne Université. Consulté le 17 mai 2024, à l'adresse https://sciences.sorbonne-universite.fr/actualites/le-stockage-des-donnees-sur-ladn-une-technologie-revolutionnaire
20. Poinar, H. N., Schwarz, C., Qi, J., Shapiro, B., MacPhee, R. D. E., Buigues, B., Tikhonov, A., Huson, D. H., Tomsho, L. P., Auch, A., Rampp, M., Miller, W., & Schuster, S. C. (2006). Metagenomics to Paleogenomics : Large-Scale Sequencing of Mammoth DNA. *Science*, *311*(5759), 392-394. Consulté le 22 mai 2024, à l'adresse https://doi.org/10.1126/science.1123360
21. Maes, A., Peillet, J. L., Julienne, A., Blachon, C., Cornille, N., Gibier, M., Arwani, E., Xu, Z., Crozet, P., & Lemaire, S. D. (2022). La révolution de l'ADN : biocompatible and biosafe DNA data storage. *bioRxiv (Cold Spring Harbor Laboratory)*. Consulté le 30 mai 2024, à l'adresse https://doi.org/10.1101/2022.08.25.505104
22. Chut. (2022, août 31). *Stockage ADN : la révolution miniature*. *Chut !* Consulté le 30 mai 2024, à l'adresse https://chut.media/tech/stockage-adn-toutes-les-donnees-du-monde-dans-une-tablette-de-chocolat/

ANNEXES

Table des annexes

ANNEXE 1 : CRITERES DE PERENNITE POUR LE CHOIX D'UN FORMAT DE DONNEES[67]

Identifiant et intitulé(s)	Définition	Justification
CPO-SOC. Communauté d'utilisateurs / Sociabilité	Le format est-il largement utilisé dans sa communauté cible ? Par le grand public ? Par les institutions de conservation ?	L'utilisation d'un format au sein de sa communauté est un indice de son adaptation aux besoins spécifiques de cette communauté. Un format également utilisé au-delà des institutions de conservation fournit des garanties supplémentaires, car les moyens de telles institutions pour maintenir un format sont limités comparés à ceux des industries culturelles. Ce critère est lié à celui des outils disponibles : plus la communauté d'utilisateurs est conséquente, plus elle est susceptible d'avoir développé ou fait développer des outils adaptés.

CPO-DOC. Documentation	Les spécifications du format sont-elles publiées ? Si oui, sont-elles maintenues par un organisme de normalisation reconnu ? Quel est leur coût ?	Si les spécifications du format sont librement accessibles, il est possible à tout un chacun d'en comprendre la structure et, si le format est également libre, de développer des outils qui le prennent en charge. La documentation peut être partielle : ainsi certains industriels publient-ils des documents décrivant seulement la structure générale de leur format. Un processus de normalisation garantit que l'on dispose de spécifications décrivant l'intégralité des caractéristiques significatives d'un format. Les formats maintenus par des organismes de normalisation nationaux (AFNOR) ou internationaux (ISO, W3C, IETF, *etc.*) offrent de meilleures garanties de maintenance et de disponibilité des spécifications, mais peuvent se révéler payants. On parle de standards *de facto* lorsqu'une spécification produite par une organisation est majoritairement adoptée dans une communauté bien qu'elle n'ait pas fait l'objet d'un processus de normalisation officiel.

[67] Caron, B. (2021). *Formats de données pour la préservation à long terme : la politique de la BnF*. https://bnf.hal.science/hal-03374030

CPO-LIB. Liberté d'utilisation	Existe-t-il des obstacles juridiques à l'utilisation du format ?	Si un format peut être totalement ouvert (documenté et utilisable par quiconque), il arrive également que des limitations d'usage pèsent sur des formats documentés, notamment en raison de brevets accordant des droits de propriété industrielle déposés au profit d'une organisation donnée. Ces brevets peuvent limiter ou interdire le développement d'outils prenant en charge le format.
CPO-AUT. Indépendance / autonomie	L'utilisation du format requiert-elle d'autres formats, encodages, environnements logiciels ou matériels ?	La consultation et l'utilisation d'un fichier numérique sont systématiquement dépendantes d'un environnement technique. Outre la dépendance à un environnement logiciel qui peut être propriétaire, abordée dans le critère « Liberté d'utilisation », l'utilisation de certains formats est tributaire d'environnements matériels, de bibliothèques logicielles, ou d'éléments habituellement non embarqués dans le fichier (par exemple, la dépendance de la plupart des PDF aux polices installées sur le poste de l'utilisateur).

CPO-ROB. Robustesse	Le format dispose-t-il de mécanismes pour repérer, ignorer voire corriger des parties altérées du signal ?	Ce critère évalue la résistance des fichiers de ce format à l'altération. Cette altération peut provenir d'une dégradation du support ou d'une erreur du matériel de lecture, mais elle est plus souvent encore le fait d'un transfert interrompu, notamment en raison d'une défaillance du réseau ou de la connectique. De ce fait, les formats de fichiers destinés à être échangés sur le réseau par le biais du *streaming* sont souvent conçus pour être robustes. La robustesse inclut les notions de résilience et de résistance à l'erreur. Elle dépend de la structure du format. Elle peut être renforcée par la présence d'empreintes numériques caractérisant chacune des zones d'un fichier, ce qui permet à un outil de validation d'identifier précisément la zone corrompue On notera que certaines méthodes de compression, particulièrement celles s'appliquant à l'ensemble du fichier et non à chacune de ses parties, peuvent avoir un effet négatif sur la robustesse des données. En raison de la réduction de la redondance que ces méthodes impliquent, une altération pourra affecter simultanément plusieurs zones du fichier.

CPO-COM. Compacité	Le format exprime-t-il une quantité d'information conséquente dans un espace contraint ? Si cette compacité est liée à une méthode de compression, celle-ci est-elle réversible (sans perte d'information) ou non ?	Un des risques majeurs pesant sur la pérennité des données est le risque budgétaire. Si les budgets que l'utilisateur peut allouer à l'achat d'espace de stockage sont limités, le critère de compacité peut devenir décisif. La compacité peut être liée à la structure d'encodage des données ou, le cas échéant, à la méthode de compression. Une compression sans perte, également dite « réversible », permet généralement une réduction significative du poids du fichier tout en garantissant la possibilité, grâce à la même méthode, de décompresser le fichier et d'obtenir une copie exacte, au bit près, du fichier source.
CPO-OUT. Disponibilité d'outils de traitement	Existe-t-il des outils de restitution, de validation, d'analyse, de migration ? L'organisme de maintenance du format en développe-t-il officiellement ?	La disponibilité d'outils de traitement est liée à la documentation (plus un format est documenté, plus les chances sont élevées que des outils de traitement l'exploitent avec précision), à la sociabilité (plus sa communauté d'utilisateurs est étendue, plus elle a de chances d'avoir développé des outils de traitement nombreux et efficaces). On prêtera une attention particulière au degré de prise en charge native par les navigateurs web comme l'indice de l'adoption du format et la garantie supplémentaire de la diffusion d'outils de lecture auprès du grand public.

CPO-ADD. Contenu additionnel embarqué	Le format permet-il d'embarquer des flux complémentaires nécessaires à l'utilisation, l'identification et la gestion du fichier (métadonnées, documentation, visuels associés, *etc.*) ?	Outre le contenu principal du fichier, le format peut être conçu pour permettre d'embarquer du contenu additionnel nécessaire à son utilisation. En l'absence de telles dispositions, l'utilisateur sera amené à transmettre séparément ce contenu, avec les risques de perte que cela implique. Le contenu additionnel peut être constitué de métadonnées permettant l'identification précise du contenu du fichier, des agents ayant contribué à sa création, des droits associés, *etc.* Il peut égalements'agir de flux spécifiques comme un visuel pour unfichier audio MP3, des sources pour un documentPDF/A-3 résultant d'une migration, *etc.*
CPO-PRO. Mécanismes de protection	Le format dispose-t-il de mécanismes de protection de son contenu ?	Ce critère est ambivalent. Certains mécanismes de protection visant à interdire l'accès ou l'utilisation d'une ou plusieurs des fonctions du fichier, tels que les DRM, peuvent empêcher l'utilisateur de mener à bien des opérations à but de conservation. Ceux à l'inverse qui n'ont pas pour but de limiter l'accès ou l'utilisation, tels les signatures électroniques embarquées comme XMLDSig pour XML, sont un atout pour garantir l'intégrité et l'authenticité du fichier.

CPO-SIM. Simplicité	Le format a-t-il une structure simple ou complexe ?	Maintenir une compétence et des outils sur un format complexe demandera nécessairement un investissement plus lourd que sur un format simple. Une méthode de compression ajoute un niveau de complexité supplémentaire. Selon les méthodes, ce niveau de complexité peut être conséquent (ex. : JPEG 2000) ou plus limité (ex. : MP3).
CPO-STA. Stabilité / évolutivité	Le format connaît-il une évolution soutenue et des versions qui se succèdent à une fréquence élevée ?	Suivre l'évolution d'un format fréquemment mis à jour peut s'avérer complexe et coûteux en investissement ; à l'inverse, un format qui ne connaît plus d'évolutions depuis plusieurs années supposera un effort moindre d'adaptation à son évolution. Plus généralement, ce critère interroge le stade de développement du format : est-il dans sa phase d'expansion initiale ou a-t-il atteint sa maturité, voire est-il toujours maintenu ?
CPO-TRA. Transparence	Le format est-il aisément lisible et compréhensible ou sa structure est-elle opaque ?	En l'absence d'outils spécifiques, un format relativement transparent pourra être plus facilement interprété et compris par un humain à l'aide d'outils génériques tels que des éditeurs de texte, XML ou hexadécimaux. La mise en œuvre d'une compression limite généralement le niveau de transparence d'un format.

ANNEXE 3 : GRILLE D'ENTRETIEN 1

Introduction de la technologie du stockage sur ADN
1. Depuis quand intéressez-vous à cette technologie ? 2. En quoi consiste cette nouvelle technologie ? 3. Comment exploitez-vous les données stockées sur ADN ? (Séquenceur) 4. Selon quelle priorité prévoyez-vous stocker les données sur les supports d'ADN ? Et pourquoi ?
Avantages et limites
1. Quels sont les principaux avantages de cette nouvelle technologie par rapport aux supports traditionnels ? 2. Quels sont les défis et les limites liés à cette technologie ? 3. Comment prévoyez-vous de les résoudre ?
Accès, intégrité et sécurité
1. Que pensez-vous du temps et du coût d'accès dédiés au stockage des données sur des capsules d'ADN ? 2. Dans quelle mesure cette technologie respecte-elle les normes de stockage d'archivage utilisées pour faciliter l'interopérabilité et la migration des données ? Selon vous, quels sont les standards qu'on devrait respecter lors du processus de conservation ? 3. Comment garantissez-vous l'intégrité et la sécurité des données encodées (mutation, délétion, suppression, *etc.*) ?
Avenir de la technologie
1. Quel est votre avis sur l'obsolescence de ce support ? 2. Quelles sont vos respectives d'avenir concernant ce support ?

ANNEXE 4 : RETRANSCRIPTION DE L'ENTRETIEN EN VISIO REALISE AVEC PIERRE CROZET

Depuis quand vous intéressez-vous à cette nouvelle technologie de stockage sur les supports d'ADN ?

Depuis, décembre 2018, pour être précis. Quand on a commencé à travailler vraiment dessus. Pour l'histoire, c'est un journal étudiant qu'il avait vu une publication, je ne sais plus laquelle, je crois que c'est celle de Microsoft, sur le stockage d'information sur le numérique sur ADN. Ils avaient stocké les archives, partie de toutes petites parties, les archives du festival jazz sur de l'ADN en partenariat avec Bio Science. Et en fait, elle a vu un article qui a été remet dans la presse et un groupe d'étudiant qui a un journal, qui s'appelle Alma Mater et un transuniversitaire sur Paris. En fait, l'un d'eux qui connaît Stéphane, ils ont demandé. « Est-ce que c'est si dur à faire ? » Donc, après avoir étudié la question, il leur dit bon. Et du coup, il a donné le bâton pour se faire battre, puisqu'ils ont dit, bah, faites-le alors ». C'est comme ça que le projet est né.

C'est là où vous avez fait les deux capsules de premier déclaration de droits des femmes et des hommes.

Stéphane a dit, ok, pourquoi pas ? Mais dans ce cas-là, trouvez-moi quelque chose d'intéressant. Et donc, ils ont réfléchi, ils ont dit, bah, on pense qu'il faudrait faire la déclaration de droits de l'homme du citoyen, il a dit oui mais il faut mettre celle de la femme aussi. C'est là où le truc a commencé.

Et en quoi consiste vraiment cette nouvelle technologie ?

En fait, ce qu'il faut bien savoir, c'est qu'on stocke des données informatiques. Le premier point, c'est que c'est clé. On ne stocke pas n'importe quoi, on stocke des données informatiques. Et donc, si on stocke des données informatiques, c'est encoder en binaire, 0 et 1. Et donc, il faut faire, c'est déjà traduire ces données binaires, en format donc de 0 et 1 en format ADN ACTG. Donc là, il faut un code, il faut une façon de coder, il y a plein de façons de faire. Nous, en fait, on a choisi de faire comme George Church, le premier fondateur du domaine en 2012 et le premier à avoir encodé de l'information numérique c'est un code qu'on appelle redondant à une base, un bit par base. En gros, quand on a un 0, on le traduit par A ou C, quand on a un 1, on le traduit par un T ou G, cette opération, on fait sur ordinateur. Et une fois qu'on a la séquence, là, il faut la synthétiser. Et donc, ce qu'on fait, la plupart du temps, c'est qu'on fait de la synthèse chimique d'ADN, avec une méthode qui s'appelle la chimie Phosphoramidite.

Elle consiste en quoi cette méthode ?

En fait, c'est une méthode de chimie qui utilise des bases qui sont produites par chimie de synthèse, en gros, par du pétrole. Et donc, on va modifier chacun des groupes des bases qui pourraient réagir de façon non contrôlée, pour pouvoir ajouter petit à petit de façon contrôlée la base qu'on veut.

Et après, dès que vous avez constitué ces petites capsules d'ADN vous les stockez comment ?

Alors avant d'arriver jusque-là, une fois qu'on a la molécule synthétisée, c'est là où on va la stocker. Dans la capsule, mais la capsule, c'est une technologie qui appartient à une entreprise qui s'appelle Imagène, à Bordeaux. C'est leur technologie.

Nous, on en a chez BIOMEMORY, on en a une autre, la DNA-Card, une autre chose qui n'a pas la même vocation. Une fois que c'est stocké dedans … bah en fait, ça ça va intéresser l'archiviste, L'ADN est sensible à trois facteurs. Premièrement, l'eau : l'ADN va s'abîmer au cours du temps, ce n'est pas un instantané, bien sûr, mais il va s'abîmer au cours du temps avec de l'eau. Il va réagir avec l'eau et il va en fait casser les molécules. Le deuxième, c'est l'air, l'oxygène dans l'air va oxyder l'ADN et donc il peut changer le message qui est encodé sur l'ADN. Le troisième facteur, c'est la lumière. C'est les rayons UV en particulier qui vont en fait casser les molécules d'ADN.

Donc là, on parle des brins d'ADN, mais on ne parle pas de la capsule ?

Non, la capsule est inventée pour la protéger l'ADN de ces trois facteurs. Et donc c'est opaque protège de la lumière, C'est dans sous atmosphère sans oxygène, que de l'argon dedans. Et l'ADN est déshydratée, on enlève l'eau de l'ADN.

Je n'ai pas compris comment vous stockez l'ADN dans les capsules ?

En fait, il faut qu'on la synthétise, ça devient quelque chose, c'est quelque chose de physique. Dans ça que je tiens dans ma main, on le prend, on le met dans la capsule on le déshydrate, c'est Imagène qui fait ça, les déshydrate, ils mettent de l'argon et ils le ferment en laser... Et ça, c'est la fameuse capsule. Et là en bas, c'est scellé au laser, complétement fermé. Et c'est comme ça, ils garantissent que la DNA peut être stable, pendant 50 000 ans.

Et après pour lire les données, stockées sur ces capsules, on utilise les séquenceurs ?

Tout à fait, indépendamment du ce sur quoi on le stocke, sur les capsules, les cartes ou papier, plusieurs façons pour stoker l'ADN. Et en fait, indépendamment de ça. Une fois qu'on en récupère l'ADN pour le lire, on le séquence. Donc il y a trois technologies de séquençage qui existent :

La première génération, c'est ce qu'on appelle séquençage Sanger, ça se fait à l'aide d'une PCR. La deuxième méthodologie, c'est ce qu'on appelle le séquençage par synthèse. Et en fait, c'est Illumina qui a inventé ça. Il y a la troisième génération et une lecture de longs fragments. Et ça c'est Oxford nanopore, en fait, on est capable de lire chaque brin d'ADN. En fait, ces technologies ont été développées pour des fins biologiques, pas des fins de lecture des données informatiques. Donc en fait, on les utilise, mais ce n'est pas forcément absolument optimal pour la lecture informatique. Mais bon. Dans tous les cas ça a été inventé pour séquencer des génomes, comme illumina, il a été développé pour ça, parce qu'on a séquencé le génome humain avec la méthodologie type de Sanger. Et ça a pris 10 ans et ça a coûté 2 milliards de dollars. Avec Illumina, ça a coûté 10 000 dollars, Donc ça avait vraiment baissé et c'était beaucoup plus rapide. Et avec oxfordnanopor aujourd'hui, il y a des gens qui annoncent 100 dollars le génome. En fait, il est très puissant car on lit chaque molécule d'ADN. Alors, Illumina, évidemment, on lit chaque molécule d'ADN mais en fait, on doit les amplifier avant, pour attendre un seuil critique, pour qu'on puisse observer à notre échelle. Alors que là, en fait, qu'on observe, c'est un courant électrique. Il y a une intelligence artificielle qui est capable d'analyser ce courant électrique et de dire qu'il correspond à tel base. C'est ça le séquençage via le Oxford nanopore. Donc, le problème qui est en général, la question que vous allez me poser après, c'est que la lecture de l'ADN est destructrice. C'est-à-dire que le brin qu'on lit on ne peut pas le récupérer.

D’accord, mais après, vous faites comment ?

Alors, ce n’est pas un problème, Si c’était un bout de papier, c'est un problème parce qu'on l’aurait uniquement. Imaginez-vous que là, on travaille sur une échelle microscopique. C'est-à-dire qu'on a ce qu'on appelle le nombre d’Avogadro pour nous, le nombre d’Avogadro, en gros, combien il y a de molécules dans une grandeur, sa valeur $6.022*10^{23}$. C'est monstrueux. Donc, même si on lit un million de molécules, il nous reste encore tellement. Donc ce n'est pas vraiment un problème. C'est comme si on avait un million de photocopies. Ce n’est pas grave et en plus, on a des méthodologies pour recopier.

c'est une méthodologie qui se base sur des bactéries, les chériechacolies ?

Oui. En fait il y en a plusieurs. Ça dépend de la molécule que vous avez. N'importe quelle molécule d’ADN, parce qu'on copie par PCR. Mais en gros, on peut tout faire par PCR. Recopier avec une bactérie. Il faut en fait qu'il y ait les signaux sur la molécule pour que la bactérie comprenne qu’elle doit le copier. Si je vous donne un texte, sans information, vous avez peut-être le lire. Mais, si je ne vous dis pas qu'il faut le copier, vous allez pas le faire. Donc, la bactérie, c'est pareil. Ça veut dire qu'on rajoute des signaux. On appelle une origine de réplication pour donner le nom exact. Il va en fait dire à la bactérie, ça c'est une molécule, il faut copier. Et donc, quand on a juste synthétisé l’information numérique et qu’on l’a mise sur ADN, on va avoir, un fragment d’ADN linéaire et qu'il n'a pas cette information. Donc, il faut la rajouter. Et en fait, c’était une des premières choses que nous l’a fait. C'est-à-dire que tous les autres, ils font juste la synthèse seule. On a rajouté une étape de ce qu'on appelle de clonage.

Après, ces bras qui contient l’information dans une molécule, on appelle un plasmide, qui contient les informations nécessaires pour que la bactérie puisse répliquer, faire une copie de cette molécule.

Mais là, après cette réplication, elle peut altérer l'intégrité des données, par exemple, faire des mutations ?

Toute copie, quelle qu'elle soit, y compris sur des matériels informatiques, ou quelqu'un qui recopie à la main, peut faire des erreurs. N'importe quel événement de copies peut faire des erreurs. Là où la biologie est excessivement forte, c'est qu'elle fait une erreur, tous les 1 milliard. Elle recopie un milliard de base, elle fera une erreur. La PCR, fait beaucoup plus. C'est un tous les 1 million. Pour vous dire, un disque dur, c'est une erreur tous les millions. En fait, les disques durs font plus d'erreurs que la biologie.

Selon quelle priorité prévoyez-vous de stocker les données sur les supports d’ADN ? Et pourquoi ?

En fait … Ce n’est pas à nous de choisir la donnée. Un vendeur de disque dur, vous ne dites pas ce que vous devez stocker. Nous on va se positionner comme un vendeur de disque dur. Ce qu’on en fera chez BIOMEMORY, on a développé un nouveau procédé. Et en fait, ce qu'on est en train de commencer à faire, est de construire une machine, qui va intégrer ce procédé. Cette machine en gros sera comme un disque dur, elle va écrire de l’ADN, elle va lire de l’ADN et stocker l’ADN dans les Card. Et donc on va se positionner comme un vendeur de disque dur. Donc, n'inverse pas nos métiers, c'est aux archivistes de savoir ce qu'on doit stocker sur le long terme. On propose un support, à l'utilisateur de savoir ce qui est important à stocker

Vos utilisateurs sont des archivistes ou des utilisateurs qui proviennent de divers domaines ?

Aujourd'hui, la majorité des données en informatique mondiale sont dans les Datacenter, donc, ça sera nos futurs clients, les gens qui ont des Datacenter. Ça peut être évidemment les GAFAM puisqu'ils ont évidemment, des Datacenter : Google, Microsoft, Amazon, Facebook, ils ont des Datacenter, qui font partie de leur business. Mais il ne faut pas croire que... Il y a plein de gens qui ont des Datacenter. Et si vous prenez quelqu'un... d'entité que vous connaissez, c'est très bien, je pense, l'INA, Ils ont le plus gros Datacenter de France. Donc, tous ces gens-là sont nos futurs clients, c'est là où on les mettra. Après, s'il y a des particuliers qui veulent stocker ça sera cher parce que c'est des infrastructures, enfin. On va faire, c'est des machines industrielles. Ce n'est pas un petit disque dur comme celui qu'on met dans notre ordinateur dans un côté de notre ordinateur. Le premier objectif, c'est les Datacenter. Et pareil, ce n'est pas eux qui choisissent ce qu'elle doit laisser en stock. Et en gros, ils vous disent, vous avez besoin de 100 Go pour 7 ans…

C'est là, en fait, le rôle de l'archiviste...

Et alors, pour moi l'archiviste il arrive en deuxième temps, sur le choix de sauvegarde des données à très long terme. D'ailleurs, on avait un petit peu parlé avec Bruno Richard, le directeur des Archives Nationales. C'est tout comme ça. J'ai le sentiment, je ne vous parle rien à sa place, mais le sentiment que j'avais, c'est clairement que l'archiviste se positionne en second temps, c'est d'ailleurs que les gens génèrent les données qui vont être stockées et les archivistes posent des questions de qu'est-ce qui vaut le coup de conserver pour le futur. Une fois qu'on a passé les textes institutionnels, ok. La déclaration de droits des hommes et des femmes, ok. Donc on a passé ça, qu'est-ce qu'on garde d'autre ? C'est là, où l'archiviste se place.

Quels sont les principaux avantages de cette nouvelle technologie par rapport aux supports traditionnels ?

En fait, les supports de l'information des supports, disque dur, SSD et bandes magnétiques, ou même DVD ont tous les mêmes trois problèmes. Un, ce n'est pas fiable, c'est-à-dire que ça ne durera pas très longtemps. Les disques durs, dans les Datacenter, ils sont remplacés tous les trois ans, voire même tous les deux ans. Les bandes magnétiques, c'est plutôt tous les sept ans. Donc, ils ne sont pas fiables. Deuxième problème, ils occupent une certaine place et ce n'est pas négligeable. Alors, le chiffre qui est souvent donné que nous on donne, en général, c'est aujourd'hui, c'est un millionième de la surface immergée du globe. Mais vu l'explosion de générations de données qu'on a, on s'attend à 2040, à ce qu'il se soit un millième de la surface c'est 10 fois plus… Mais il faut que ça commence avec pas du tout négligeable. Et si on veut amener les informations sur Mars, on ne peut pas en amener deux fois Paris. Aujourd'hui, les Datacenter, c'est deux fois Paris. Le troisième problème, c'est que ces supports, ils consomment de l'énergie. Il faut fonctionner certes, mais surtout pour les conserver. Il doit être conservé à une certaine température… En fait, c'est écrit. Vous prenez un truc de banque magnétique, c'est écrit dessus. À conserver, entre 16 et 22 degrés. Et entre 35 et 60 d'humidité. Donc, les Datacenter, il y a des climatiseurs monstrueux. Ça pose des questions écologiques.

Donc, les supports sont obsolescents. Ce qui m'amène aux avantages de l'ADN, quand il est bien conservé, comme ça, c'est 50 000 ans avec une température ambiante. Et ensuite, ça, c'est pour l'énergie nécessaire pour la conservation et la

stabilité. Et en plus, là, dans ma main, c'est l'équivalent d'un grain de sable, moins qu'un grain de sable. Et c'est 100 milliards de copies. En gros, il pourrait faire tenir toutes les données du monde dans une boîte à chaussures. Pour tout, ça ne demande pas d'énergie. Une fois que c'est stocké dessus, l'écriture, la lecture, demande de l'énergie, évidemment. Mais le stockage, non.

Et dernière chose, l'ADN n'est pas obsolète, tant qu'il y a des biologistes et des médecins, on aura besoin de savoir manipuler de l'ADN, le support ne sera pas obsolescent…

Quels sont les défis et les limites liés à cette technologie ?

Il y a deux défis qui résoudraient les deux limites de l'ADN. C'est qu'en fait, on essaie de faire chez BIOMEMORY, en fait, synthétiser l'ADN, c'est long et cher. Donc c'est les deux défis auquel on s'attaque. Justement. La lecture peut être lente aussi, mais pour l'instant, c'est loin d'être limitant. Ça pourrait être long si l'on considère qu'on va atteindre un niveau d'un SSD mais pour l'instant, on est encore là. J'aurai pour résoudre les écrits. Donc l'écriture en gros, ça dépend qui vous dit, si chez TWICE BioScience, c'est 1000 $ de mégaoctet. Il faut arriver au moins à un 1 $ le Go. Ça a un énorme impact. On pense qu'on est capable de faire chez BIOMEMORY.

Pour la rapidité, ça c'est plus une question de procédé. Par exemple, on a des nouveaux brevets c'est pour aussi pouvoir aller plus. On va faire comme l'informatique. Ce n'est pas un processeur qui va plus vite. C'est mettre un million de processeurs en parallèle C'est la parallélisation. C'est ce qu'on fait. Notre procédé il a les capacités d'être paralysé. On travaille de la parallélisation et miniaturisation et diminuer les coûts aussi. On a travaillé par les matières premières. On n'utilise pas de briques qui viennent du pétrole. On utilise ceux qui viennent du vivant, ils sont beaucoup moins cher à obtenir. Et en plus en écologie, c'est une des idées.

Que pensez-vous du temps et du coût d'accès dédié au stockage des données sur les captures d'ADN ?

Il faut plus indépendamment de la capsule. Parce que la capsule, c'est une technologie qui a été conçue pour être un coffre-fort. Donc, logiquement, pour ouvrir un coffre-fort, ça prend du temps. C'est pour ça que nous on a changé le support. Donc je ne critique pas ~~à~~ ce support-là (capsule). Si je veux un coffre-fort, ça c'est très bien. Si je veux quelque chose qui stocke mes données un certain temps, peut-être pas 50 millions mais qui permet d'accéder plus rapidement, on ne prendrait plus celui-là (card). Donc, le temps d'accès, c'est variable. Ça dépend de ce qu'on veut faire. L'idée, c'est d'arriver à accéder comme l'informatique. Essayer d'accéder au premier octet avec le temps le plus faible possible. On se comparant à ce qu'il est comparable, c'est-à-dire les bandes magnétiques par exemple.

Eh oui, et en ce qui concerne le coût ?

Le coût, notre objectif qu'on a, c'est un dollar le téraoctet en 2030 chez BIOMEMORY.

Dans quelle mesure cette technologie respecte-t-elle, les normes de stockage d'archivage utilisées pour faciliter, par exemple la migration des données ?

En général, la migration des données se fait parce que le support devient obsolète et que la raison de son obsolescence. Les bandes magnétiques, tous les sept ans, on recopie tout, Là, je l'ai dit 50 000 ans. Donc dans la L'ADN card, en gros 150 ans sans problème. Donc le problème ne se pose pas exactement de la même

façon. Donc, elles sont la migration des données. Et les données sont écrites sur deux supports différents. Il peut y avoir un événement qui va toucher un type de support et qui n'affectera pas un autre. Donc, en fait, le plus simple, c'est de relire la donnée et de l'écrire sur quelque chose d'autre. Donc en fait, comme nous, de toute façon, notre solution finale va intégrer de la lecture, pour le recopier sur un autre support. Une fois, ça se fait au départ, soit après, c'est-à-dire qu'on écrit sur ADN, on stocke et si on veut le mettre sur notre support, on relie, on recopie, par exemple, on veut y conserver sur ADN et après, une fois qu'on a, on les stocke sur un disque dur, voilà, il va être jamais migré en fait. Dans une logique archiviste, le plus simple, c'est d'avoir stocké sur ADN à deux endroits différents.

Et en ce qui concerne la sécurité et l'intégrité ?

Nous on stocke l'ADN. Donc on a déjà vérifié qu'il n'y a pas de mutation. Il y a qu'on fait un contrôle qualité après chaque synthèse. Pour être sûr qu'on a la bonne molécule. Et après, une fois qu'on la protège de l'eau, de la lumière et de l'air, de l'eau ça se passe rien.

Et concernant la sécurité, tout ce qui est stocké sur les capsules d'ADN ?

Elles sont quand même encodées d'une certaine façon. Il faut quand même savoir comment elles sont encodées. Donc, ce n'est pas forcément évident. On a besoin de logiciels, ils sont propriétaires à BIOMEMORY, ils ne sont pas sur la place publique. Aujourd'hui, quand on parle de sécurité, 99 % des questions sont reliées à la cyber-sécurité. Si on va accéder à la donnée, il faut l'avoir physiquement, comme un livre. Donc, l'ADN, ça peut en plus, c'est microscopique. On peut avancer de plein de façons de faire pour le cacher. C'est vraiment où on veut. On ne veut que ce soit pas évident. L'ADN on peut le cacher dans plein de choses. Il y a un sujet qui peut être vous intéresser à qui il s'appelle, le L'ADN of things…Comme par exemple, quand je cachais de l'information, si je prends dans cette feuille de papier, sur laquelle je mets de l'ADN, je le sèche et ça, je peux le garder 20 ans, sans trop de problème. Si vous savez pas qu'il y a l'ADN sur cette feuille de papier, bah ..

Après, on peut aussi crypter de plein de façons. L'information, on peut la crypter et de façon informatique. Ça, on n'a pas encore vraiment commencé à travailler dessus. On a un projet dont je ne peux pas parler, mais on fait un peu ce genre de choses. Il y a trop de choses à faire au niveau de sécurité en plus que les traditionnels.

Quel sont vos perspectives concernant l'avenir de ce support ?

Je vais prendre les mots de Montecillo, l'ADN il ouvre des perspectives faramineuses. Pour le stockage, les données certes, comme l'exemple d'aller sur Mars, on peut avoir toutes les données pour aller sur Mars. Mais il ouvre autre chose. En fait, il ouvre une ère de façon de voir L'ADN, ça ouvre une ère de ce qu'on appelle l'informatique moléculaire, C'est bien d'avoir les données sur L'ADN, c'est mieux de pouvoir faire quelque chose de ces données sur ADN. Et ça, c'est ce qu'on appelle le biocomputing, donc on peut faire des opérations, comme dans un ordinateur, avec l'ADN.

On va commencer à explorer chez BIOMEMORY, pas instantanément, on va commencer, on a déjà commencé à réfléchir. Il y a pas mal de gens qui ont déjà commencé à réfléchir et différentes façons de faire. On peut le faire avec les molécules L'ADN. En fait, on peut aussi le faire en utilisant des cellules, c'est-à-dire encoder les problèmes informatiques et donner une des cellules et c'est elles qui sont problèmes. Parce qu'en fait, la façon dont l'on a eu d'encoder fait qu'il y a des

signaux que la bactérie est capable de comprendre et donc elle va traiter l'information.

Comme la méthode de duplication pour les bactéries ?

Oui, c'est ça. C'est utiliser les bactéries pour faire quelque chose. Et en fait, il y a déjà des gens qui ont ça en biologie de synthèse. Par exemple, on peut apprendre à des bactéries à jouer au morpion. Et à ce qu'on peut apprendre en plus, donc ça, c'est des circuits génétiques. Le circuit qui est derrière, c'est des opérateurs booléens. On appelle aussi des opérateurs logiques, le ET et OU. Alors, et ça, en fait, c'est la base de l'électronique et la base de l'informatique.… Après c'est les data manager qui s'en occupent des Datacenter, et de préciser de stocker à tel endroit parce que c'est de la donnée chaude ou tiède ou froide ou ultra froide mais ça c'est le qui j'ai arrondi mais il y a six mois je vous aurais dit les données froides mais en fait ce n'est pas exactement ça, oui c'est une chose qu'on vise mais potentiellement on peut viser type de données à part peut-être les ultra chaudes, en moyen terme mais il y a des technologies qui vont sans doute encore le développement qui pourront changer la lecture et d'aller plus vite encore.

ANNEXE 5 : GRILLE D'ENTRETIEN 2

La technologie du stockage aux Archives Nationales
1. Quels sont les supports utilisés au Archives Nationales pour la conservation des données numériques ? 2. Pourriez-vous me préciser la quantité des données numériques conservées aux Archives Nationales ?
La technologie du stockage sur ADN
1. Quelles sont les nouvelles technologies de stockage des données numériques ? 2. Que pensez-vous de la nouvelle technologie de stockage sur des capsules d'ADN ? 3. Avez-vous participé à l'événement d'encodage sur deux capsules d'ADN la Déclaration des droits de l'homme et du citoyen et la Déclaration des droits de la femme et de la citoyenne ?
Archiviste face à la nouvelle technologie d'ADN
1. Quel était le rôle des archivistes lors de cet événement ? 2. Avez-vous remarquer l'absence des écrits des archivistes dans ce bouleversement qui affecte leur son cœur métier ? 3. Selon vous, le supports d'ADN présentent-ils une option fiable pour pérenniser des données numériques ?

ANNEXE 6 : RETRANSCRIPTION DE L'ENTRETIEN TELEPHONIQUE REALISE AVEC VAN DE WALLE THOMAS

Quels sont les supports utilisés aux archives nationales pour la conservation des données numériques ?

Alors, on utilise des serveurs. Donc avec des technologies très classiques d'aujourd'hui, on fait de la conservation de ces données sur serveur. On ne fait pas de la conservation sur le support froid.

On fait, par contre, une copie sur des supports froids type t sur le banc de LTO, on trouve la conservation de froid incite. Le type LTO, c'est un type de support de stockage (des disques magnétiques), c'est comme les disques durs.

Quelle est la quantité de données numériques conservées aux archives nationales ?

Alors, il y a deux types de données numériques conservées aux archives nationales. Il y a les données numériques qui sont des copies de documents papier. Voilà. Donc là, on a une vingtaine de Téra. Mais ça fait longtemps que je n'ai pas consulté ces chiffres. On appelle les archives numériques du coup. J'ai une hésitation parce qu'on les conserve en différents exemplaires en fonction de besoin. Et ensuite, il y a tout ce qui est données numériques natives. C'est-à-dire des archives qui n'ont jamais existé sous forme papier, qu'on récupère sous format numérique, et ça, ces archives-là, c'est 130 téraoctets.

Avez-vous des idées sur les nouveaux supports de stockage ?

Oui, oui, oui. On se tient à peu près au courant. Après, si vous voulez notre problématique, nous, c'est on est des utilisateurs de ces supports du stockage, on travaille avec d'autres services du numérique, du ministère de culture, qui sont en gros responsables de tous les choix informatiques du ministère. Donc, c'est eux qui déterminent les choix technologiques de stockage, et on apporte nos besoins au métier. On dit qu'on voudrait du stockage pour faire ça, et c'est eux qui déterminent des solutions ensuite techniques.

Nous, notre sujet par rapport à ça, c'est d'avoir des solutions techniques qui correspondent à nos besoins, mais qui correspondent aussi aux problématiques d'exploitation facile par les gens qui les gèrent, et les gens qui les gèrent, c'est les agents du service du ministère des cultures. Donc les technologies qu'on utilise, sont des technologies mûres et prouvées, industrialisées, si vous voulez, pour nos usages au quotidien. Après, on fait de la veille sur les technologies, et on fait des actions ponctuelles comme stockage sur ADN, mais ce sont des actions ponctuelles de recherche et d'innovation.

Et donc, que pensez-vous de la nouvelle technologie de stockage sur les capsules d'ADN ?

C'est ce que je viens de vous dire. C'est une solution, et en fait, de nombreux intérêts, en termes de capacités de stockage, l'aspect écologique, l'aspect de consommation. Voilà. Mais bon, son grand défaut, ce n'est pas une solution facilement utilisable à une échelle industrielle du quotidien.

Donc voilà, nous, l'intérêt qu'on a eu de travailler sur l'écriture sur ADN de la Déclaration des droits de l'homme et du citoyen et la Déclaration des droits de la femme et de la citoyenne, c'est d'avoir de participer à de la recherche organisée par l'université, en lien avec des sociétés privées émergeantes. Voilà, pour démontrer l'intérêt de cette technologie, mais ce n'est pas un choix. Et par ailleurs, aujourd'hui, on sait que ça peut marcher. On sait qu'on peut relire parce que c'est ce que la société nous dit. Mais il faut ré accéder à l'information facilement, car on est obligés de passer par des tiers, par des intermédiaires. Et puis la technologie d'écriture de fabrication de ces bras d'ADN, c'est encore des choses qui prennent du temps.

Vous avez participé à l'événement, d'encodage sur les deux capsules d'ADN, les deux déclarations ?

Alors, moi, j'étais associé d'assez loin à ce projet. J'étais, à ce moment-là, responsable du projet Adamant, j'ai suivi à distance de mon côté. J'en ai suivi les résultats avec un grand intérêt, mais je n'étais pas partie prenante.

Mais selon ce que j'ai compris, les archivistes, ils ont fourni seulement les documents nécessaires pour l'encoder. Et puis après, on a participé d'un point de vue symbolique, en prenant en charge les capsules encodées avec ces nouvelles technologies et d'assurer que ces capsules sont conservées avec les originaux. Voilà. Donc, il y a un côté... Oui, des manifestations de notre part, d'un soutien à cette innovation, et notamment parce que c'est aussi porté par une entreprise française et avec la recherche française. Donc, on est dans cette logique d'action publique, au sens large et de volonté de soutenir la recherche française d'une certaine manière.

Avez-vous remarqué l'absence des écrits des archivistes sur ce sujet, qui affectent leurs cœur de métier ?

Alors, je n'ai pas eu cette curiosité, mais ça ne m'étonne pas... Ce n'est pas, parce qu'on fait pour les archivistes, la solution est stockage, comme son nom l'indique, c'est que la solution c'est qu'un support, ce qui nous importe nous c'est l'information. Le support, c'est une solution technique qui est définie avec des spécialistes des supports par rapport à ce qui nous a besoin pour les archivistes. Et les besoins portés par les archivistes, c'est des besoins de conservation et des besoins d'accès à l'information. Et si vous voulez, à partir du moment où on fait constater rapidement que l'ADN pour les archivistes, c'est un sujet du futur… Là, aujourd'hui, ce qui nous préoccupe, comment est-ce qu'on gère les informations qui nous sont confiées, comment est-ce qu'on gère les métadonnées qui permettent d'accéder aux informations, comment les traiter plus efficacement, comment on arrive à prendre en charge au grand quantité de l'information. Et les solutions qui nous permettent de stocker les volumes, donc vous parliez tout à l'heure, c'est-à-dire 130 téraoctets, on n'a pas besoin vraiment de l'ADN. Et puis, par ailleurs, voilà, pour écrire sur l'ADN, s'il faut des mois, on reçoit nous des archives toutes les semaines. Donc si, chaque fois, pour écrire des archives, chaque semaine, il faut attendre 1 mois pour que ce soit écrit, si vous voulez, ce n'est pas efficace. Voilà. Donc ce n'est pas efficace. Les préoccupations des archivistes ne sont pas centrées sur les solutions de stockage, c'est un sujet parmi X, mais voilà, c'est finalement relativement secondaire quand même. On nous a parlé d'autres solutions, sur diamant qui offrent d'autres possibilités, et du stockage sur et avec des solutions électromagnétiques.

Enfin, les solutions de stockages technologiques sont très nombreuses. Nous, ce qu'on a besoin, c'est de solutions fiables, qui nous permettent de ré accéder à l'information, et qui peuvent être exploitées facilement par les gens qui travaillent au quotidien, et sur lesquelles on peut facilement écrire et lire.

Selon vous, le support d'ADN présente-t-il une option fiable pour pérenniser des données numériques ?

Un jour… Oui.

Et en ce qui concerne les données froides ?

Oui, oui, tout à fait, oui, non, quand on vise un jour, je voulais dire vraiment par rapport à nos usages. Comme je vous dis d'aujourd'hui, si vous voulez nous aux Archives Nationales, on conserve 130 téraoctets. Donc, pour stocker sur des supports froids, on utilise les bandes LTO pour une durée de vie de 5 à 10 ans et on les recopie ensuite. Pour l'instant, est-ce qu'utiliser de l'ADN est mieux que ces solutions-là ? Ce n'est pas évident par rapport aux investissements qu'il faut faire pour aller dans ces directions. Et sachant que disque LTO, on peut le tester assez régulièrement, l'ADN, de ce que j'ai compris une fois qu'on rouvre la capsule, qu'on déshydrate et qu'on voit des codes pour vérifier que toutes les informations sont encore là. Si on peut faire ces vérifications, est-ce que j'ai compris, mais on a perdu notre support ?

Oui. C'est ça.

Ça veut dire qu'il faut aussi avoir une stratégie de production de capsules d'ADN liée aux vérifications régulières qu'on souhaite faire, par rapport à l'intégrité de l'information où on fait complètement confiance à la technologie en disant qu'on met ça dans un coin et on est sûr. On fait qu'on est sûr que dans 100 ans, 200 ans, 500 ans, 10 000 ans, les informations sont toujours là sans changement.

En tout cas, pour conserver des données sur support à froid, il y a des technologies de stockage concurrentes. Et qui vu les volumes qu'on traite sont encore dans le paysage. On n'a pas d'obligation à aller vers l'ADN, la variété des solutions de stockage à froid, parce qu'après d'un point de vue énergétique, par définition, le stockage à froid, ça ne consomme rien après l'écriture.

Le stockage à froid, si on parle vraiment du stockage à froid, vous enregistrez sur un disque dur ou dans une bande LTO et de le mettre sur un étager. Ce n'est pas de laisser dans un robot. Si vous avez fait votre gravure sur un LTO et vous la mettez sur une étagère, ça ne consomme plus rien. La laisser sur un serveur ou la laisser dans un robot, alors on est plus sur du stockage à froid, on est sur un stockage à chaud.

Si on parle vraiment de stockage à froid, c'est dans les caractéristiques de l'ADN, on stocke de l'information, pendant X temps, on n'inquiète plus, et je ne sais pas besoin d'accéder pendant un temps. Il y a d'autres solutions aujourd'hui, alors qui sont moins stables dans le temps…

.. Il y a les données stockées à froid et il y a les données non utilisées, mais qu'on veut garder utilisables qui sont conservées à chaud dans des serveurs, à partir du moment où c'est des serveurs, c'est à priori à chaud. Si vous les gardez sur des serveurs, c'est que vous voulez avoir la possibilité d'y accéder à tout le monde. Et en fait, si vous voulez mettre à froid, c'est que vous partez du principe que vous acceptez un délai pour réaccéder à l'information. Et que donc, il faut passer soit par un robot qui va vous cherchez une bande quelque part, soit vous passez par un humain qui va aller chercher la bande, c'est-à-dire la recharger sur un serveur. Et seulement à ce moment-là, vous y aurez accès.

Et si je traduis avec la L'ADN, la seule solution pour y accéder et la dernière, c'est du froid! c'est l'AD, où on peut pas y accéder pour une demande immédiate … Parce qu'en fait, le stockage à froid, c'est des solutions de sécurisation d'information

pour éviter plusieurs choses, pour éviter notamment les problématiques de risque de perte, à cause d'une panne, à cause d'une attaque informatique, etc. On fait des copies à froid et on les met dans un coin. Voilà, on se dit, si je perds mes copies à chaud, mes informations à chaud ou que je perds mes serveurs, j'ai toujours mes données dans un coin.

Donc l'ADN pour ça, c'est intéressant par rapport à cette optique, mais par rapport à la quantité de données gérées dans le monde conservées sur serveurs …

Est-ce que les grandes entreprises Amazon et Google, ils ont des stratégies de stockage à froid de conservation de l'information, si oui quelles informations ? Alors là, voilà, il faut investir en plus, il faut faire des copies, il faut gérer des copies, il faut gérer la possibilité de réaccéder à des copies, on parle d'une branche très particulière du stockage de l'information. En général, c'est pour la sauvegarde et pour la sécurisation.

En ce qui concerne, BIOMEMORY, ils travaillent pour le moment leur défi. C'est ces deux axes-là, c'est le temps et le coût… et leur futur client seront les grandes entreprises, qu'ils ont des gros serveurs comme Amazon, Facebook, Google.

Oui, dans une logique de préservation de l'information, mais bon, quand on fait du stockage à froid pour sécuriser des données, il faut tenir compte des problématiques de flux.

C'est-à-dire que si vous avez, c'est vraiment par rapport à la population de données, vous voulez conserver au sein de toutes les données que vous avez. Si l'idée, c'est de sauver des données et de stocker des données à tout moment, la problématique, c'est que le temps que vous écriviez vos données combien de nouvelles en arrivent…

ANNEXE 7 : GRILLE D'ENTRETIEN 3

La technologie du stockage au CC-IN2P3
1. Quelle est la typologie les données stockées au sein du CC-IN2P3 ? 2. Sur quel support sont stockées ? 3. Pourriez-vous me préciser la quantité des données numériques conservées dans les serveurs du CC-IN2P3 ?
La technologie du stockage sur ADN
1. Selon vous quelles sont les nouvelles technologies de stockage des données numériques ? 2. Que pensez-vous de la nouvelle technologie de stockage sur des capsules d'ADN ? 3. Vous pensez que le centre peut l'utiliser comme support à la place des serveurs ? 4. Selon vous, le supports d'ADN présentent-ils une option fiable pour pérenniser des données numériques ou pour créer des copies de sauvegarde?

ANNEXE 8 : RETRANSCRIPTION DES EXTRAITS DE L'ENTRETIEN REALISE AVEC JEAN-YVES NIEF

« … J'étais dans une conférence sur le stockage de Mars il y a deux semaines. L'avenir du stockage, les technologies, les supports, qui seront utilisés à l'avenir, et ça a encore très éloigné. On va reparler un petit à l'heure, mais pour l'ADN, ça semble encore très éloigné. Mais en moins, c'est des choses qui regardent vraiment très loin, il y a des peut-être des choses qui sont plus proches pour nous qui sont par exemple, justement, il y avait une présentation comme ça. le stockage sur céramique, c'est de la lithographie, de base et de gravures. Il y a un laser avec des perspectives on y est beaucoup plus proches et facilement adaptables pour un centre de données, ça nous amène quand même à l'échelle de la prochaine décennie, à partir de 2030. Mais peut-être un horizon plus proche par rapport au stockage sur ADN.

Je vous rappelle l'institution, on est le centre de calcul à la base de l'institution du CCIN2P3.

Le CCIN2P3 est un acte de recherche fondamental du CNRS. On est de façon primaire ont fourni des moyens de calculer de stockage pour les expériences de physique fondamentale essentiellement dans les domaines que je viens de dire en plus de l'astroparticule, car on mélange l'astrophysique de physique de particule ou des technologies de détection, et aussi une ouverture l'hébergement des serveurs pour les humanités numériques, et aussi, on a des conventions avec d'autre institut comme INE, qui est l'institut national de l'écologie, des écosystèmes, et de l'environnement pour laquelle on fournit des services. Alors, typiquement, ce dont les chercheurs ont besoin, ils produisent énormément de données, d'observations de détecteurs et de simulation numériques, et ces données-là, forcément, sont tellement volumineuses et à l'heure nous avons environ 250 pétaoctets de données, qui sont stockées chez le centre.

Donc, c'est un vol conséquent et forcément, qu'on se fait de façon automatisée. On ne veut pas dire qu'il n'y a pas de présence d'être humain, bien évidemment, il faut des chercheurs, des ingénieurs, pour pouvoir développer les codes informatiques, pour pouvoir gérer les analyses de données, dans les tâches de calculs, gérer les données qui sont stockées chez nous. Ça, c'est du travail, mais la tâche, on a même de voir les données, de pouvoir les triturer, c'est à la base un gros travail qui fait mon niveau informatique. On a des algorithmes classiques, et puis maintenant, un petit peu beaucoup de l'intelligence artificielle. Donc, des choses qui sont très, très gourmandes.

En infrastructures, il y a l'infrastructure, des serveurs des calculs, du stockage, du disque, etc.

Et on a des besoins divers et variés. On est au niveau du stockage pour de la donnée non structurée, c'est-à-dire, le fichier c'est très souvent les données de l'expérience de physique, des fichiers, qui sont indépendants les uns des autres, qui correspondent à une séquence de ce qui doit être observé. Aussi, on a aussi des gens de pouvoir les faire en ce moment, les index, etc. avec des connaissances métiers, c'est-à-dire beaucoup de bases de données, les volumétries sont beaucoup plus faibles, mais l'utilisation de bases de données est extrêmement importante, parce qu'elles permettent justement de référencer ~~et~~ des données et de pouvoir les rechercher plus facilement. Donc, tel type de données, ils étaient en telle condition.

Je peux les sélectionner à partir de catalogue des bases de données.

Les bases de données qui peuvent être réellement essentielles et qui peuvent être dans certains cas très conséquentes aussi notamment on va héberger dans une base de données qui correspond à des observations de base pro physique, donc c'est l'SSD, qui est un télescope qui est basé à Ochili, où le CC IN2P3 va contribuer à échelle de 30 ou 40 % pour l'aspect ressources de calcul et de ce coupage et là, donc, ils vont faire entre 3 nuits un balayage complet du ciel avec une caméra qui fait 1 milliard de pixels. Donc, ça veut dire 15 téraoctets de données brutes.

On a plusieurs aspects au niveau du stockage pour les gens, on laisse payer cet aspect de catalogue donc avec besoin d'un accès extrêmement rapide. Je mets à part le L'SSD, on utilise un besoin d'accès très rapide pour être très performant, pour avoir des réponses qui sont immédiates et pour ça, on utilise en technologie des bases de données relationnelles et derrière tout ça, au niveau du média physique, ce qui nous intéresse plus aujourd'hui, des disques flashes donc, on est très loin de l'ADN, bien évidemment, parce qu'on a besoin, on peut avoir besoin...

Ce qu'on a c'est des centaines de mégaoctets par seconde, et aussi, ce qu'on appelle de l'accès aléatoire c'est-à-dire qu'on ne va pas, quand on dit, des données sont base de données, on n'a pas A à Z, mais on va picorer à droite à gauche, on va aller à un endroit, après un autre endroit donc, d'autres parties de l'espace physique du disque, on a aussi les disques mécaniques, les disques qui tournent, ou on a l'habitude, c'est déjà un petit peu limite. Donc, c'est pour ça qu'on se dirige. On s'est dirigés vers tout le plus dans le disque le SSD, qui nous permet justement d'avoir des temps d'accès beaucoup rapides et de faire plein d'opérations par seconde. Donc, on a une petite quantité de stockage qui représente pas grande chose en volumétrie par rapport au reste, et qui nous permet de faire jusqu'à 1,5 millions d'opérations d'entrées sorties par seconde. Parce qu'il est au-dessus de ce que nous avons besoin, mais qui permet, même, de pouvoir se revenir aux besoins.

« … Et surtout, derrière le temps c'est très bien C'est bon, l'ordre de moins de la mini seconde. C'est-à-dire que l'opération sur une base peut durer au moins d'une demi seconde. On a besoin de ça pour justement avoir un temps de réponse raisonnable parce que, généralement, ces bases de données sont en fait les cerveaux des systèmes d'information qu'ils sont. »

« .. Donc, on a cet aspect-là, on a encore beaucoup de disques mécaniques, parce que ce n'est pas cher pour le gros de la volumétrie pour tout ce qui est données non structurées. Parce que finalement, dans le contexte actuel, ça convient ça convient en termes de performance, parce qu'on est capable en multipliant le nombre de disques de pouvoir soutenir nos besoins. Typiquement, à la l'heure actuelle on a au bon, allez 600 serveurs pour le stockage, sur des disques mécaniques, avec un peu de SSD à l'intérieur, mais pas grand-chose, et donc, ça fait des dizaines de milliers de disques pour des données non structurées, qui est vraiment la donnée la plus... plus populaire, en ce moment, qui est plus utilisée chez vous. Ce qui permet, en ce cas, de l'avantage de la donnée non structurée, par rapport à la donnée structurée, comme dans une base de données, ça permet de pouvoir finalement faire de facilement des du traitement parallèle. C'est-à-dire qu'on a une grosse masse de données. »

« .. En fait, je vais diviser ma tâche en plein de petites tâches qui vont chacune lire un certain nombre de fichiers. On peut faire tout ça en parallèle, et on a l'aspect important aussi à chez nous. C'était là que, bon, on est 250 pétaoctets de données je ne vais pas oublier que toutes ces données sont précieuses. Et on doit vous besoin

d'être accédés aussi à la santé tout le temps. Il y a des contraintes, tout le niveau économique. En deux types de contraintes, tout le niveau économique, c'est que nous avons des risques mécaniques, c'est certes, il est devenu assez démocratisé. Mais il n'est pas... Enfin, comme on a des grosses volumétries, parce qu'on a beaucoup d'achats, chaque année, on n'a pas des achats pour plusieurs centaines de milliers d'euros, pour pouvoir renouveler du l'ancien matériel ça, c'est un problème aussi tous les 5 ans ou les 7 ans, et on fait une augmentation de capacité parce que les besoins aussi des utilisateurs augmentent. C'est l'aspect énergétique aussi, qui devient extrêmement important. Et aujourd'hui, c'est-à-dire que, on a des disques qui tournent, qui tournent pour ne rien faire, même si on ne l'est pas de données les disques tournent, les disques mécaniques, et qui consomment beaucoup d'énergie. Nous, on a un centre de calcul que sur la puissance, 1.4 mégawatt on avait une facture d'électricité de l'ordre de 1,2 ou 1.3 millions d'euros par an. Donc, l'année dernière 2023, avec la crise au niveau du marché de l'énergie, on s'est retrouvé avec une facture de plus de 3 millions d'euros. Eh bien nous, on a un budget constant, c'était il y a qu'on a quand même...On a des contraintes importantes au niveau budgétaire, mais on doit pouvoir répondre aux besoins des gens de façon la plus économique aussi. Alors, donc, c'est pour ça qu'il y a un aspect absolument pas par exemple pour l'instant et qu'est-ce qui est chez nous et qui est vraiment profondément fait dans la culture de nos centres de données et qui prend beaucoup d'ampleur ailleurs que soit dans privé ou dans public, et les clients privés à Amazon, etc. Donc, c'est le stockage sur bandes magnétiques qui est extrêmement important, c'est-à-dire que là, c'est un avantage énorme. Et puis, en plus de ça, au niveau économique, c'est-à-dire qu'au niveau énergétique aussi, c'est-à-dire que de la donnée qui est sur une bande magnétique, elle est dans une bandothèque robotisée. Donc, elle est appelée à la demande, c'est-à-dire que le robot prend la bande magnétique et la met dans un lecteur pour la lire de la transférer sur le disque, et puis c'est-à-dire par des utilisateurs, tout ça de façon transparent pour l'utilisateur. Mais du coup, pas de la donnée qui n'est pas nécessairement lu en permanence, et bah on la laisse sur la bande magnétique, c'est beaucoup moins cher. Pour celles, il y a quand même quelque chose d'essentiel. Et surtout, au niveau des performances, en lecture écriture, des avantages qui sont... C'est-à-dire qu'on a des bits, en lecture écriture 300 et 400 mégaoctets par seconde. Ce qui est en phase avec d'autres besoins, et avec la volumétrie. »

C'est des bandes magnétiques comme les bandes LTO, c'est ça ?

« Oui, donc, on a de la LTO pour la partie sauvegarde. Pour la partie vraiment l'essentiel, on est sur la bande propriétaire du jaguar, de chez IBM. Mais IBM fait partie du constructeur des LTO. Disons, on se peut dire, c'est que le jaguar est un petit peu en avance toujours par rapport aux performances d'LTO. Toujours une petite avance de génération. Là, à l'heure actuelle, par exemple, on est sur la LTO 9, avec des capacités de 18 téraoctets par bande qui sont sur le jaguar, on est sur la dernière génération qui est apparue récemment de 50 téraoctets. Ça veut dire que la prochaine génération la LTO dix, soit aux alentours de la centaine de téraoctets. On va agir de façon très particulière, parce que la plupart du temps, les bandes magnétiques, c'est vraiment pour les données froides. Nous, on l'utilise vraiment comme un système de stockage qui peut être accédé à tout moment. Donc, ce qui met des contraints très importantes sur le système. Pour vous donner un ordre d'idée à l'heure actuelle, on peut descendre jusqu'à 3 téraoctets par jour de données, la bande magnétique verra les disques pour la lecture. Donc, on a besoin de beaucoup de performance. »

« … C'est ça seulement pour la lecture. Mais en termes énergétiques ça consomme moins que quand un serveur. des bandothèques, à l'heure actuelle, on a des grandes bandothèques c'est de l'ordre de quelques kilowatts, la puissance qui est demandée, donc c'est l'équivalent de 3-4 radiateurs, il y a un simple serveur avec une bande, vous voyez, c'est un gros avantage. Tout en ayant, la flexibilité de pouvoir redescendre les données à la demande. Nous, ce qu'on fait .. Ce n'est pas du offline, c'est-à-dire, ce n'est pas des lignes. Les données n'ont pas par ligne. Il ne faut pas une action humaine pour mettre la bande sur des étagères. On les met sur des lecteurs, tout ça. Elles sont accessibles à tout moment, à tout moment, des journées de la nuit. C'est des contraintes technologiques pour nous Ça veut dire, on dépend énormément des bandothèques parce que c'est bien beaucoup de mécaniques. Voilà, ça peut créer des problèmes pour un système qui est dans un moins stable en production, qui marche bien, et qui donne satisfaction, mais qui rajoute de complexité en niveau ingénierie Mais c'est comme ça. C'est le plus à payer pour pouvoir avoir accès à un stockage très accessible, performant et économique… »

« … Alors, donc, on a deux types de sauvegarde. On a un service de sauvegarde, parce que, voilà, que lors de quelques pétaoctets de données, 3.5 pour être précis actuellement. Donc, qui sert pour sauvegarder les données précieuses des bases de données, notamment pour aller récupérer. Donc, ils sont des bases de données, des configurations de serveurs, informatiques aussi, donc, si le serveur casse, si on va récupérer la configuration des serveurs on peut faire la récupération au quotidien. Voilà, ça, c'est un aspect. Et là, on utilise la LTO, parce qu'on ne va pas avoir besoin de beaucoup d'accès, etc. Mais, là, on y a les données qui sont sur bandes magnétiques pour pouvoir... Bon, pour la donnée scientifique, qui est très grosse volumétrie, voilà, parce que tout dépend des politiques, qui sont mises en place sur les collaborations scientifiques.

On est loin de l'archivage, on est loin de la sauvegarde, mais ça permet, c'est... c'est un une façon fait. Mais non, on est des plus petits projets entre les domaines, mais qui sont des choix des collaborations à travers la semaine. Là, souvent des collaborations internationales, pour lesquelles nous sommes un unique centre de traitement de données, et puis, en plus, comme là, on pose dans ces caractes, c'est faire double copie sur bandes magnétiques. Parce que ça nous arrive, vraiment, mais une bande magnétique peut casser, avec quelques incidents par an, pas beaucoup, mais il suffit qu'il y ait un lecteur qui ne soit pas en forme et qu'il casse la bande, donc, on a un certain niveau de protection qui va dépendre de tout le monde. Il faut que tout le monde en ait conscience, et bien sûr que, soit, de côté des administrateurs service, et des utilisateurs, on a un certain niveau de sécurité, qui est ce qu'il est, et qu'il faut accepter.

Par rapport à l'avenir, la bande magnétique a clairement encore plus d'ampleur, on est sur un rythme de densification qui est beaucoup, beaucoup plus important, ou plus rapide, que pour du disque mécanique ou du disque SSD. Parce que, tout simplement, sur ces derniers médias, on se rapproche, de l'échelle atomique, car on est à l'échelle du nanomètre du milliardième de mètre. Donc, c'est de plus en plus compliqué, des données sur le réduire, de le commencer. On est ce que, sur la bande magnétique, on a l'avantage d'être sur des densités qui sont beaucoup plus faibles. Néanmoins, ça vient avec des inconvénients.

La conservation des bandes magnétiques, sur les dernières générations qui sont en train de sortir et en ce moment-là, on se retrouve avec un nouveau rejet. L'un dernier moment, et ça veut dire que ça devient même s'il est beaucoup plus

intrinsèquement, il est beaucoup plus fiable que du disque solide, du SSD, ou du disque le disque mécanique. »

« L'autre aspect aussi, l'âge de 30 ans, 40 ans, ce qui est bien bien, nous on a des siècles de vie plus longs que les bandes magnétiques que sur le disque, c'est 5 ans, 7 ans, et il faut bien le reposer. La bande, on est sur des cycles de 8 ans, 10 ans. Donc ce qui est beaucoup mieux, on peut dire pourquoi vous êtes censés, vous vous dites, 30 ans pour la durée de vie d'une bande et 10 ans le problème, c'est que même si les constructeurs garantissent leur média pour une durée de 30 ans. Actuellement, les médias deviennent obsolètes au bout de 10 ans, 15 ans, même de la bande magnétique. Mais il y a un aspect important, y compris pour de l'archivage de données, on a des gros volumes qui sont archivés, on va être capables de les écrire, et justement, là, à l'heure actuelle, c'est que le stockage sur ADN est extrêmement intéressant.

Pour ce sens en termes de densité, c'est l'équivalent d'un centre de calcul de sa machine de 800 m^2 c'est peut-être une petite capsule, mais le gros, gros, gros souci actuellement, c'est le coût et le temps d'accès. Voilà, c'est ça, c'est l'accès. On est à 100 kilobits par seconde vous écrivez par jour 10 gigaoctet, donc, on doit être capables d'écrire de l'ordre de 400 ou 500 téraoctets par jour. C'est-à-dire de l'ordre de 40-50 gigabits et plus par seconde. Mais encore un peu, et bientôt on va arriver à l'échelle pétaoctet par jour. Donc, il y a des défis de se lancer dans le stockage sur ADN. Donc, c'est pour ça que c'est hors de portée pour l'instant.

Et je ne sais pas à quelle échelle de temps on pourra compter sur stockage sur ADN. On est intéressés clairement, on ne va pas être les premiers utilisateurs de stockage sur ADN. Parce que, ce qui nous intéresse, c'est faire de beaucoup de lectures et écritures et ce genre de choses. Donc, pour l'archivage je pense dans le domaine des bibliothèques, tout ça, ça peut être beaucoup plus intéressant sur du plus court terme. Et pour nous, ça ne l'est pas pour l'instant. Mais même pour faire de l'archivage massif comme une bibliothèque nationale, etc.

Y a quand même le souci effectivement du débit. On a pour nous dire, on a des performances qui sont verraux de façon exponentielle mais puis l'apparition des premiers disques durs dans les années 50. C'est-à-dire qu'en fait, dans côté les médias, on m'a aussi beaucoup plus rapidement peu les capacités au niveau réseau connectique pour pouvoir accéder aux données. Donc, on a parlé de pression parce que voilà, en fait, on multiplie les médias au niveau du nombre de serveurs, etc. On peut pouvoir accéder de la même façon au goulier. Une façon efficace, mais en fait, notre capacité à accéder un octet d'informations, les performances ont baissé depuis les années 60. Ce qui peut paraître étonnant. Mais on a quand même ce décalage, on a cette progression ou un rapide des performances au niveau capacité réseau.

Mais c'est compensé par la multiplication des serveurs. Mais là, par exemple, si on pense de stockage ADN donc, notre problème, il est... là, il est à des échelles au-dessus. Donc, côté, on a un stockage qui est extrêmement dense. Voilà, donc pour en résumer pour nous, ce qui... technologie,

je m'y serai près de pas là-dessus dans les médias, alors je parlais de la céramique tout à l'heure, donc ça, ça pourrait être plus intéressant, c'est que la céramique par exemple dans un rac, donc vraiment dans une armoire, il est leur objectif pour 2030, c'est pouvoir mettre une centaine de pétaoctets avec des débits de l'ordre de quelques gigas par seconde on sait un peu moins bon, même moins bon que ce qu'on est capable de faire avec disque magnétique voire de la bande magnétique mais ce n'est déjà pas mal. Donc c'est plus évident. »

Selon vous, les supports d'ADN présentent-ils une option fiable pour pérenniser des données numériques ou pour créer des copies de sauvegarde?

Je dirais pour l'archivage, parce que la sauvegarde, en fait, comme c'est essayer de faire la récupération, c'est quelque chose qui est une sauvegarde. Donc, une sauvegarde de serveur, c'est pour faire la récupération en une catastrophe. Donc c'est des données qui changent tout le temps. Sur l'archivage, c'est vraiment de la donnée qui était documentée qui bouge plus. Oui. Effectivement, ça pourrait être. On pourrait avoir, façon en archivage en cas classique sur 3 copies, avoir des copies sur des technologies différentes, des médias différents, et sur des sites éloignés, des autres. Des sites distincts, ça peut être de sa machine différents. Effectivement, des solutions pour être d'avoir une copie, une des copies sur ADN c'est clairement quelque chose de très intéressant.

TABLE DES MATIERES