

Élaboration des spécificités d'un connecteur ADN pour un archivage conforme au modèle OAIS

**Travail de master réalisé par :
Thomas KERBOUL**

**Sous la direction de :
Pierre-Yves Burgi, PhD**

Genève, 15 août 2024

**Information science
Haute École de Gestion de Genève (HEG-GE)**

Déclaration

Ce travail de Master est réalisé dans le cadre du Master en Sciences de l'information de la Haute école de gestion de Genève.

L'étudiant-e atteste que le travail rendu est le fruit de sa réflexion personnelle, a été rédigé de manière autonome sans avoir utilisé des sources autres que celles citées dans la bibliographie et a été vérifié par un logiciel de détection de plagiat.

L'étudiant-e accepte, le cas échéant, la clause de confidentialité.

L'utilisation des conclusions et recommandations formulées dans ce travail, sans préjuger de leur valeur, n'engage ni la responsabilité de l'étudiant-e, ni celle du-de la directeur-trice.

Fait à Genève, le 15 août 2024

Thomas Kerboul

Remerciements

C'est avec plaisir que je remercie mon directeur de recherche, Pierre-Yves Burgi, pour son accompagnement et ses encouragements tout au long de ce travail. Les multiples séances de suivi ont été l'occasion d'échanger sur les possibilités de l'archivage sur ADN au-delà du cadre du mémoire.

Mes remerciements vont aussi à mes collègues du master en Sciences de l'Information pour leur bonne humeur et l'entraide réciproque durant ces trois années d'études.

Résumé

L'explosion des volumes et l'obsolescence rapide des technologies de stockage actuelles poussent au développement de nouvelles solutions pour répondre à l'enjeu de la préservation à long terme des données numériques. Dans ce contexte, l'acide désoxyribonucléique (ADN) émerge comme une solution prometteuse pour relever ces défis en raison de son extrême densité, de sa stabilité dans le temps, de ses faibles coûts énergétiques et de son absence d'obsolescence. Toutefois, stocker des informations à long terme n'est qu'une première étape : pour garantir leur accès dans un futur lointain, il est essentiel d'intégrer l'ADN dans un cadre archivistique tel que celui fourni par le modèle OAIS (Open Archival Information System), utilisé depuis plus de deux décennies par les professionnels de la préservation numérique.

Situé à la croisée de deux disciplines rarement combinées, ce mémoire définit les spécificités nécessaires à l'élaboration d'un connecteur ADN conforme au modèle OAIS tout en prenant en compte les contraintes propres aux particularités de ce support de stockage. Grâce à une revue approfondie de la littérature secondaire, ce mémoire s'ancre dans les cadres théorique et pratique de la préservation numérique et tire parti des dernières innovations relatives au stockage sur ADN.

Le cœur de ce mémoire est composé de recommandations pratiques pour l'implémentation du connecteur ADN, selon deux scénarios aux caractéristiques différentes tant dans les formes que prend ce support de stockage que sur l'intégrité structurelle des fichiers archivés. Une documentation au format analogique devant nécessairement accompagner une archive ADN pour permettre la récursivité de l'information de représentation selon le modèle OAIS, nous formulons tout d'abord un certain nombre de recommandations sur son contenu et sur sa forme.

Puis, pour le scénario simplifié, nous proposons une solution basée sur la concaténation des amorces et sur l'existence de plusieurs espaces de noms afin de créer une forme rudimentaire d'indexation et de transmettre certaines informations importantes dès la lecture des nucléotides. Cette approche a également été étendue à toutes les métadonnées accessibles hors d'un fichier archivé qui profitent ainsi d'un codec simplifié permettant la transmission de l'information sans passer par le binaire. Pour le scénario avancé, nous proposons deux solutions basées sur des nanostructures afin d'offrir une plus grande rapidité d'accès à l'information et une meilleure densité par rapport au scénario simplifié : une première solution utilise des QR-codes et la seconde des formes particulières basées sur des origamis ADN ou sur la topologie.

L'implémentation de ces recommandations permettra de garantir l'accessibilité du contenu des archives stockées sur ADN malgré l'inévitable évolution des technologies et de la société.

Mots clés : modèle OAIS, préservation numérique, stockage de données sur ADN

Table des matières

Déclaration.....	i
Remerciements	ii
Résumé	iii
Liste des tableaux	vi
Liste des figures.....	vi
1. Introduction.....	1
1.1 Problématique	1
1.2 Contexte.....	2
1.3 Plan	2
2. Stockage sur ADN	3
2.1 Problématiques du stockage sur supports actuels	3
2.2 Avantages du stockage sur ADN	4
2.3 Historique	5
2.4 Limites	6
2.5 Concept	6
3. Le modèle OAIS	9
3.1 Historique et justification	9
3.2 Concept	10
3.2.1 Principes généraux	10
3.2.2 Environnement	10
3.2.3 Entités fonctionnelles	10
3.2.4 Information	11
3.3 Entité fonctionnelle Stockage	14
4. Revue de la littérature sur la préservation à long terme	17
4.1 Introduction	17
4.2 Préservation de la partie immédiatement perceptible	17
4.2.1 Introduction	17
4.2.2 Libellés	18
4.2.3 Documentation	18
4.3 Préservation du support.....	22
4.3.1 Introduction	22
4.3.2 ADN	23
4.4 Préservation des données numériques.....	26
4.4.1 Introduction	26
4.4.2 Émulation	27
4.4.3 Machine virtuelle	29
4.4.4 ADN	33

5. Élaboration des spécificités d'un connecteur ADN	35
5.1 Introduction	35
5.2 Définition des concepts	35
5.3 Technologies et scénarios envisagés par DNAMIC.....	36
5.4 Spécificités communes aux deux scénarios.....	37
5.4.1 Généralités	37
5.4.2 Documentation analogique nécessaire	37
5.4.3 Documentation analogique supplémentaire	40
5.4.4 Choix du codec	40
5.4.5 Format des données	41
5.5 Spécificités pour le scénario simplifié.....	42
5.5.1 Transmission de l'amorce générique et du codec	42
5.5.2 Gestion des amorces	43
5.5.3 Gestion des métadonnées	45
5.5.4 Gestion de l'émulation	46
5.5.5 Résumé	46
5.6 Spécificités pour le scénario avancé.....	47
5.6.1 Option 1 : utilisation de QR-codes	48
5.6.2 Option 2 : utilisation de nanostructures avec des formes distinctes	48
6. Conclusion	50
Bibliographie	52

Liste des tableaux

Tableau 1. Classification proposée par la NEA pour constituer le SER	21
--	----

Liste des figures

Figure 1. Chronologie des principaux travaux publiés sur le stockage de données numériques sur ADN.	6
Figure 2. Lien entre information et données dans le modèle OAIS	12
Figure 3. Schéma des différents composants du Paquet d'informations	14
Figure 4. Schéma de l'entité fonctionnelle Stockage	15
Figure 5. Interprétations possibles d'une suite de 8 bits	26
Figure 6. Schéma des prérequis nécessaires à l'émulation	28
Figure 7. Mécanisme général de l'archivage des données selon R. A. Lorie	30
Figure 8. Étapes d'encodage de la solution Micr'Olonys	32
Figure 9. Étapes de décodage de la solution Micr'Olonys	32

« Alors ils m'ont dit qu'à l'appui des théories et des technologies les plus avancées dans tous les domaines, après un grand nombre de recherches théoriques et expérimentales, après d'interminables analyses, synthèses et études comparatives, ils étaient enfin parvenus à mettre au point une méthode qui permettrait de conserver des informations pendant une période de cent millions d'années. Ils m'ont assuré que c'était la seule technique qu'ils estimaient possible à ce jour. Cette méthode consistait à... [...] graver des mots dans la pierre. » Liu Cixin, La Mort immortelle, Babel : 2021, p. 786-787.

1. Introduction

1.1 Problématique

La masse toujours plus importante de données à stocker et l'obsolescence rapide des supports de stockage de l'information numérique poussent au développement de nouvelles technologies permettant d'assurer une densité de stockage très élevée et fiable à très long terme, sur plusieurs centaines voire milliers d'années. Depuis plusieurs années, l'acide désoxyribonucléique (ADN) apparaît comme une solution crédible, bien qu'émergente, à cette problématique, en raison de sa grande stabilité dans le temps, de sa densité de stockage plusieurs fois supérieures aux systèmes de stockage par bande magnétique et par mémoire flash ainsi que de l'absence d'obsolescence en raison de son caractère naturel et organique.

Toutefois, si l'ADN permet d'assurer un support physique stable, il ne garantit en rien l'accessibilité sémantique de l'information. Il peut être opportun de préciser d'emblée que le stockage n'est qu'une composante de l'archivage : stocker des informations à long terme, même sur plusieurs centaines d'année, n'en fait pas *de facto* des archives sans la mise en place des mécanismes nécessaires à leur bonne interprétation. Comme l'écrit un rapport publié par la Nuclear Energy Agency de l'OCDE sur la transmission intergénérationnelle de la mémoire des déchets radioactifs, « it is not just a question of handing down a message, but of keeping that message interpretable, meaningful, credible and usable over time » (Nuclear Energy Agency 2019a, p. 13). Naturellement, cela n'est pas propre à l'ADN, ni aux données numériques en général – il suffit de songer aux hiéroglyphes ou au linéaire B, déchiffré uniquement au début des années 1950, sans parler du linéaire A toujours indéchiffré, pour mesurer tout l'enjeu de la transmission d'information à long terme – mais ce défi est exacerbé par la superposition de couches à décrypter avant d'arriver à l'information.

La communauté professionnelle de la préservation numérique à long terme utilise depuis deux décennies le modèle Open Archival Information System (OAIS), publié en tant que norme internationale ISO 14721. Ce modèle définit les composants de base d'un système de préservation à long terme, les acteurs internes et externes de ce système et les fonctions de chacun, ainsi que la nature des documents archivés. Crucialement, ce modèle se contente de fournir un cadre d'application et se veut agnostique vis-à-vis de toute architecture ou d'implémentation spécifique, laissant aux organisations souhaitant l'implémenter toute latitude sur le choix des technologies.

Ainsi, le projet européen DNA Microfactory for Autonomous Archiving (DNAMIC) développe un connecteur pour archiver les données sur ADN, dont les spécificités se veulent conformes au modèle OAIS. Le but du présent mémoire est de définir ces spécificités en prenant en compte les contraintes propres aux particularités du stockage sur ADN. En particulier, ce mémoire investigate les possibilités d'auto-description des données et de prise en charge à

long terme de plan de reprise d'activité (*disaster recovery*). En d'autres termes, ce mémoire pose la question de comment garantir l'interprétabilité du contenu des archives stockées sur ADN alors que les technologies et la société auront évoluées tout au long de la durée de vie du support de stockage.

1.2 Contexte

Le présent mémoire s'insère dans la continuité de deux projets menés par l'Université de Genève et plus particulièrement par le département « Recherche et Information Scientifique » de la division du Système de l'Information (STIC), dirigé par Pierre-Yves Burgi.

Le projet Data Life-Cycle Management (DLCM) est un projet suisse visant à développer des services, destinés principalement aux chercheurs des hautes écoles helvétiques, pour la conservation à long terme des données de la recherche et des publications. Lancé en août 2015 dans le cadre du programme national « Information Scientifique », il s'est achevé en 2021 avec plusieurs livrables (Burgi, Makhoul Shabou 2021). Parmi ceux-ci, une solution pour un système de préservation des données à long terme a été développée, avec une architecture modulaire, conforme au modèle OAIS, aux principes FAIR et aux normes internationales. Deux implémentations en ont été déployées : yareta.unige.ch et olos.swiss. Au sein de cette solution, le module « Archival storage » dispose de connecteurs à différents systèmes de stockage : un système de fichier, le protocole Amazon S3 et des bandes magnétiques.

Le présent travail de recherche s'inscrit en outre dans le cadre du projet européen DNA Microfactory for Autonomous Archiving (DNAMIC), lancé en 2023 et coordonné par plusieurs hautes écoles et entreprises en Europe. Ce projet vise à développer des micro-usines mobiles, modulaires et automatisées, capables de prendre en charge toutes les étapes de l'archivage sur ADN sans aucune intervention externe, de manière à ce que toute institution puisse bénéficier d'une unité de stockage à long terme (DNAMIC 2024). Au sein de ce projet, l'Université de Genève est particulièrement impliquée dans la création d'un connecteur qui permettrait aux données d'une solution DLCM d'être stockée dans des brins d'ADN (Burgi 2024, p. 11). Il est bien admis, toutefois, que les caractéristiques actuelles du support de stockage sur ADN, que nous présenterons dans le chapitre suivant, ne permettent pas pour l'heure d'envisager le stockage sur ADN autrement que comme une copie de sauvegarde pour les documents patrimoniaux ou comme un système de soutien à un plan de reprise d'activité (Burgi 2024, p. 11).

Dans le contexte helvétique, ce travail a pu profiter de quelques recherches préliminaires : il s'agit d'une part du travail de bachelor de D. Andriamahady (2021) et d'autre part d'une présentation faite à l'IPRES 2022 (Burgi et al. 2022).

1.3 Plan

Le présent mémoire s'articule autour de plusieurs chapitres. Le chapitre 2 revient brièvement sur les problématiques actuelles de l'archivage ou plutôt du stockage des contenus numériques, avant d'introduire le rôle de l'ADN dans ce contexte. Il en détaille les avantages, l'historique de la recherche et les modalités de fonctionnement.

Le chapitre 3 introduit le modèle OAIS. Il revient sur son historique et sa justification ainsi que sur les différents concepts nécessaires à sa compréhension (environnement, entités fonctionnelles, information). Il se conclut par une analyse détaillée de l'entité fonctionnelle Stockage. Nous avons pris le parti de nous concentrer sur les points pertinents du modèle

OAIS pour le présent mémoire et de laisser de côté certains aspects. Ceci entraîne nécessairement une certaine simplification du modèle.

Le chapitre 4 présente une revue de la littérature des problématiques couvertes par le mémoire, à savoir les mécanismes de préservation à long terme et les stratégies de plan de reprise d'activité qu'il est possible de mettre en œuvre. Comme la lecture le montrera, la problématique de ce mémoire fait écho à des considérations dans divers domaines et secteurs d'activité. Cette revue de la littérature est organisée autour de trois axes : la préservation à long terme de la partie immédiatement perceptible de l'archive, la préservation à long terme du support et la préservation à long terme des données.

Enfin, le chapitre 5 constitue le cœur de ce mémoire. Il détaille les particularités et attentes propres aux projets DLCM et DNAMIC et propose des recommandations pour les spécificités d'un connecteur ADN, selon deux scénarios aux caractéristiques différentes tant dans les formes que prend le stockage ADN que sur l'intégrité structurelle des fichiers archivés.

2. Stockage sur ADN

2.1 Problématiques du stockage sur supports actuels

L'impact de la transition numérique a été universel. Tout type de secteur d'activité, tout individu, tout format d'information est concerné (Lavoie 2014, p. 4). Dès les années 1990, une majorité de l'information est sous forme numérique (Hilbert, López 2011). Une estimation prévoit que le volume de données créé en 2025 atteindra 175 zettabytes (175×10^{21}) (Reinsel, Gantz, Rydning 2018) ; d'autres projettent que la demande de stockage atteindra en 2040 entre 3 yottabytes (3×10^{24}) et 70 ronnabytes (70×10^{27}) (Zhirnov et al. 2016). Bien que seule une fraction de cet ensemble ait vocation à être conservée, son stockage, particulièrement à long terme, est une problématique à part entière. En effet, les supports de stockage actuels – entendus comme les supports de stockage développés ces dernières décennies et actuellement commercialisés – pour les données numériques sont soumis à plusieurs contraintes, dont la conjonction péjore la possibilité de réaliser un stockage à long terme.

D'une part, la densité des supports de stockage actuels atteint ses limites physique (Evans et al. 2012). Pour stocker les 175 zettabytes évoqués précédemment sur DVD, il en faudrait de quoi faire 222 fois le tour de la Terre (Reinsel, Gantz, Rydning 2018, p. 7). Les bandes magnétiques, qui ont la densité la plus élevée parmi les supports de stockage actuels, se compteraient également par millions pour stocker une telle quantité (Nguyen et al. 2020, p. 1).

D'autre part, les supports de stockage actuels sont peu stables dans le temps. Contrairement au papier, dont le stockage dans de bonnes conditions peut lui permettre de durer plusieurs centaines d'années, on estime que les supports de données numériques ont une durée théorique entre plusieurs dizaines d'années et 150 ans au maximum (Lunt 2012; Grass et al. 2015, p. 2552; Anžel, Heider, Hattab 2021, pp. 4908-4909; Wang et al. 2024, p. 3). Des observations anecdotiques soulignent toutefois que cette durée est limitée aux supports avec la meilleure qualité, sous peine d'être drastiquement diminuée à dix années au maximum (LaBarca 2012, p. 139; Extance 2016, p. 23). Quoi qu'il en soit, une telle limite n'a jamais été prouvée ; en effet, les technologies les plus anciennes encore utilisées, telles que les bandes magnétiques, n'ont pas encore 75 ans (Anžel, Heider, Hattab 2021, p. 4907).

Cet état de fait souligne un autre problème : les avancées technologiques rapides de ces dernières décennies ont créé un problème d'obsolescence des supports de stockage. Certains cessent d'être commercialisés et utilisés et les données encodées, si elles peuvent être encore viables, ne peuvent être lues faute du matériel suffisant. Afin de garantir l'accès aux données, il est donc nécessaire de procéder à de coûteuses migrations, qui ne sont pas sans risques pour l'intégrité des données. En pratique, tant pour des questions liées à l'intégrité du support qu'à la question de l'obsolescence, la durée de vie d'un support de stockage excède rarement dix ans avant que son contenu ne soit transféré sur un autre support (LaBarca 2012, p. 138; Extance 2016, p. 23).

De plus, les technologies actuelles sont confrontées à la crise énergétique et environnementale. En effet, les technologies actuelles nécessitent d'importantes quantités d'énergies pour fonctionner (Hormann, Campbell 2014; Extance 2016, p. 22). Les progrès technologiques constants ne permettent que de partiellement contrebalancer la quantité de données à stocker. On estime la consommation totale des *data centers* à 200 TWh, soit 1% de la production mondiale ; leurs émissions de gaz à effet de serre représentent quant à elles

0.3% du total mondial (Jones 2018; Masanet et al. 2020). Quant à la production de silicium, nécessaire pour la mémoire flash, elle est polluante et particulièrement gourmande en eau et en énergie (Panda et al. 2018, p. 2). Il existe également un problème d'accès aux matériaux : l'étude de Zhirnov et al. (2016) prévoit ainsi que la demande de stockage numérique en 2040 en mémoire flash nécessitera des quantités de silicium dix fois supérieures à la production mondiale actuelle.

Ces différents points soulignent que les technologies actuelles de stockage sont insuffisantes pour garantir un stockage de données en masse sur le long terme et avec un minimum de manipulation.

2.2 Avantages du stockage sur ADN

Dans ce contexte, l'acide désoxyribonucléique (ADN) est envisagé comme une alternative viable pour répondre aux problématiques des supports de stockage. L'ADN est une molécule présente dans les cellules qui constitue l'élément fondamental de la vie de tous les organismes connus. Il contient les informations génétiques nécessaires au développement, au fonctionnement, à la croissance et à la reproduction des organismes vivants. Structuellement, l'ADN se compose de deux longs brins formant une double hélice, chaque brin étant constitué de nucléotides. Ces nucléotides contiennent quatre bases différentes : l'adénine (A), la thymine (T), la cytosine (C) et la guanine (G). La disposition de ces bases le long des brins d'ADN forme le code génétique, qui détermine les caractéristiques et les traits d'un organisme.

Dans une récente revue de la littérature, Wang et al. (2024) résument les principaux avantages de l'ADN dans leur application pour le stockage d'information. Tout d'abord, l'ADN propose une densité théorique de stockage très élevée. Avec une densité de 2 bits par nucléotide et une seule copie de chaque brin, la densité pourrait atteindre environ 455 exabytes (455×10^{18}) par gramme (Church, Gao, Kosuri 2012). Toutefois, des limites pratiques tant dans la densité d'encodage des bits que dans le nombre de copies nécessaires pour permettre la récupération des données, cette densité doit se limiter à 17×10^{18} bytes par gramme, ce qui demeure respectivement six et trois ordres de grandeur de plus que les stockages par bande magnétique et par mémoire flash (Organick et al. 2020, p. 5; Anžel, Heider, Hattab 2021, p. 4905). Avec cette densité, les 175 zettabytes évoqués précédemment pourraient donc théoriquement tenir dans 10 kilogrammes d'ADN (Wang et al. 2024, p. 1). Les données encodées sur ADN peuvent en outre être très facilement et très fidèlement répliquées par une réaction en chaîne par polymérase (PCR) (Ceze, Nivala, Strauss 2019, p. 456), malgré certaines limitations biologiques (voir section 2.5).

Une autre caractéristique majeure de l'ADN est sa stabilité dans le temps. Dans des conditions optimales, il est théorisé que l'information stockée sur ADN puisse être décodable dans deux millions d'années (Grass et al. 2015, p. 2555). Plusieurs cas de figure font état d'ADN décodé avec succès sur des échantillons vieux de plusieurs centaines de milliers d'années (Panda et al. 2018, p. 3). Dans un environnement moins idéal, il resterait possible d'accéder à l'information dans 2000 ans (Grass et al. 2015, p. 2555), ce qui reste remarquablement plus élevé que la durée de vie prévue des supports numériques. La demi-vie de l'ADN à température ambiante est évaluée à plus de cent ans (Zhirnov et al. 2016, p. 376). Ces caractéristiques peuvent rapprocher l'ADN du papier ; dans la formule de Rosenthal et al. (2013, p. 514), « [p]aper as the medium for the world's memory has one great advantage; it survives benign neglect well ».

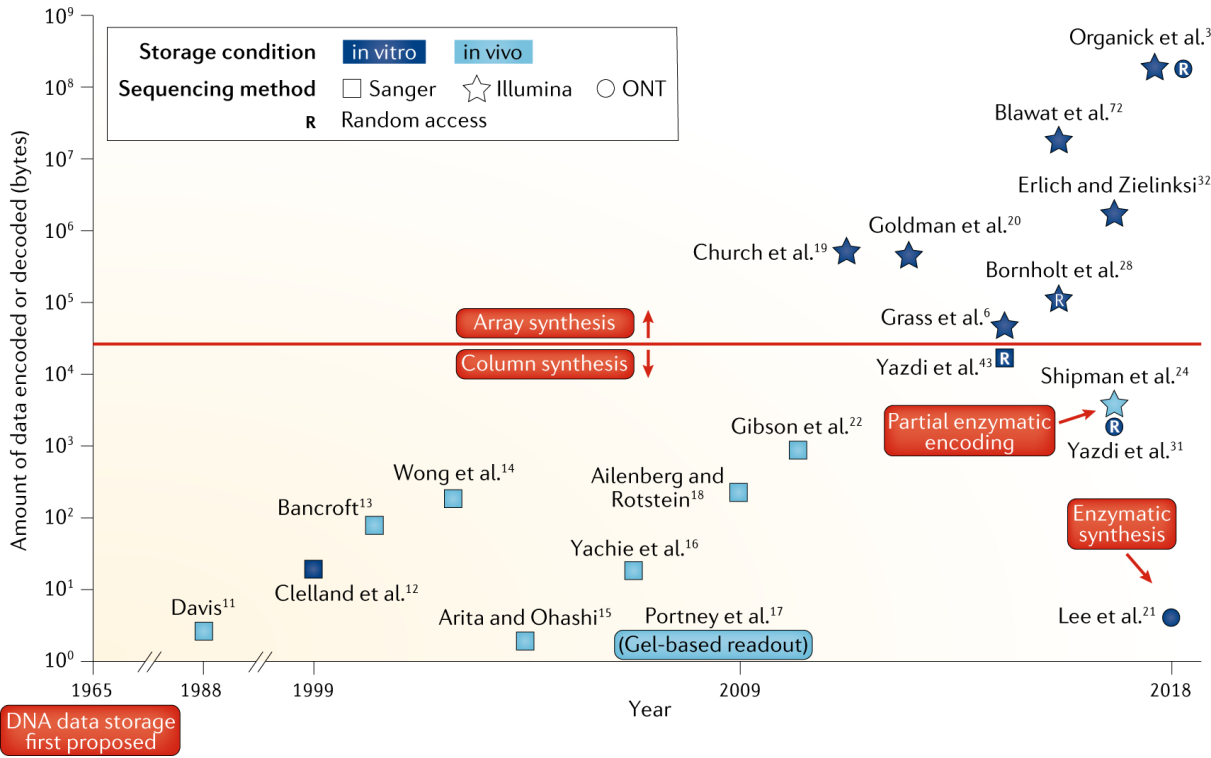
Cette stabilité dans le temps est de plus possible à un coût énergétique relativement faible, jusqu'à trois ordres de magnitude inférieurs par rapport aux bandes magnétiques (Zhirnov et al. 2016, p. 377; Extance 2016, p. 22; Panda et al. 2018, p. 4). Une étude a également montré qu'avec certaines méthodes de synthèse, le stockage d'un terabyte est bien moins polluant sur ADN que sur bande magnétique et surtout sur disques durs (Nguyen et al. 2020).

Enfin, en tant que base de la vie actuelle sur Terre, l'ADN est un support qui ne risque pas de devenir obsolète. Sans rentrer dans les considérations sur l'âge des plus anciens organismes utilisant l'ADN, on peut s'attendre à ce que des formes de vie avec ADN soient toujours existantes d'ici plusieurs dizaines de milliers d'années : les techniques de séquençage ADN, si elles peuvent évoluer, resteront donc toujours d'actualité (Ceze, Nivala, Strauss 2019, p. 456). Cette caractéristique de l'ADN le place à part dans les supports de stockage, à long terme ou non, qui sont dépendants d'une technologie spécifique pour être lu ; si D. S. H. Rosenthal et avant lui J. Rothenberg mettent particulièrement en garde contre l'obsolescence des supports de stockage à long terme, cette critique ne nous semble pas pouvoir s'appliquer à l'ADN (Rothenberg 1998, p. 2, fig. 2 et photo 1; Rosenthal 2017, p. 18). Il faut toutefois noter que cette résistance à l'obsolescence n'est vraie que tant que les brins d'ADN sont composés des quatre bases organiques que nous avons mentionnées plus haut. Si des solutions utilisant des bases composites (Anavy et al. 2019) ou de nouvelles bases artificielles (Choi et al. 2019; Ren et al. 2022), voire d'autres polymères que l'ADN (Rutten et al. 2018), conservent les avantages que nous avons précédemment décrits et permettraient même une meilleure densité, le fait qu'elles soient sans parallèle organique les met au risque de ne plus être comprises par les générations futures.

2.3 Historique

Ces différentes caractéristiques font de l'ADN une alternative intéressante pour le stockage à long terme. En effet, l'ADN en tant que support d'information miniature est évoqué dès le milieu des années 1960 par le physicien soviétique M. S. Neiman (1964). Toutefois, ce n'est que deux décennies plus tard, en 1988, qu'une expérience artistique, *Microvenus*, parvient à encoder 35 bits dans de l'ADN *in vivo* (Davis 1996) ; il faut attendre dix ans de plus pour qu'une expérience de stéganographie développe une solution de stockage sur ADN *in vitro* (Clelland, Risca, Bancroft 1999). Depuis, les expérimentations se sont succédées et un volume de plus en plus important d'information a pu être stocké sur ADN, mais ce sont surtout les travaux parallèles de Church, Gao et Kosuri (2012) et ceux de Goldman et al. (2013) qui sont largement crédités pour avoir relancé l'intérêt pour le stockage sur ADN. Le record est actuellement détenu par Organick et al. (2018), qui sont parvenus à stocker 200 mégabytes (2×10^8).

Figure 1. Chronologie des principaux travaux publiés sur le stockage de données numériques sur ADN.



(Ceze, Nivala, Strauss 2019, fig. 1)

2.4 Limites

Le stockage sur ADN, dans sa technologie actuelle, présente toutefois certaines limites, contre lesquelles D. S. H. Rosenthal met en garde (Rosenthal 2017). D'une part, l'ADN présente un important temps de latence entre la demande d'accès au document et son accès effectif ; cela est dû au fait qu'il s'agit d'un media hors ligne, mais cela est également aggravé par le séquençage nécessaire pour récupérer les données (voir section 2.5). Ce temps de latence élevé limite l'applicabilité du stockage ADN à un support de sauvegarde (Rosenthal 2017, p. 18).

De plus, si les prix du séquençage de l'ADN se sont effondrés, la synthèse représente toujours un coût significatif qui rend actuellement impraticable la commercialisation à large échelle de l'ADN (Rosenthal 2017, pp. 25-26; Panda et al. 2018, p. 7; Wang et al. 2024, p. 14). Malgré des réductions importantes des coûts ces dernières années et même en prenant en compte la facilité de répllication une fois l'écriture effectuée, les prix doivent continuer à baisser pour permettre un usage plus large.

2.5 Concept

Concrètement, le principe de stocker de l'information sur ADN passe par six étapes successives (Wang et al. 2024, pp. 2-5) :

1. Encodage : cette étape consiste à convertir les données numériques en formats compatibles avec l'ADN. Elle implique l'utilisation d'algorithmes pour transcrire les informations binaires en séquences de nucléotides ou en motifs structurés spécifiques. L'enjeu principal de cette étape est donc de maximiser l'efficacité de l'encodage afin que plus d'information

puisse être stockée par nucléotide. Une revue des différents codecs (codeur-décodeur) est faite par Garafutdinov et al. (2022).

2. Écriture : cette étape se réfère au passage des données encodées au support de stockage ADN. Trois grandes problématiques dominent cette étape : la précision, la vitesse, y compris la capacité de produire plusieurs brins en parallèle, et le coût. On peut distinguer deux grands principes, dont le choix va influencer les modalités des étapes successives :
 - a. Le stockage de données par séquence d'ADN encode directement les informations dans des suites de nucléotides, dont chacun représente une partie des données binaires. Généralement, il s'agit d'ADN à un seul brin (ssDNA). Les séquences sont assemblées par processus chimiques ou enzymatiques. Cette technique profite des technologies de lecture et d'accès de l'ADN organique, dont elle imite le fonctionnement (Wang et al. 2024, pp. 5-7). Théoriquement, chaque base pourrait stocker deux bits (A = 00, C = 01, G = 10, T = 11). Toutefois, certaines propriétés de l'ADN compliquent ce principe : ainsi, les séquences répétitives d'une même base (homopolymères) ne sont pas possibles, de même qu'un déséquilibre dans la proportion de G et de C avec les autres bases. De plus, afin de pallier des erreurs d'écriture ou de lectures, les données sont accompagnées d'algorithmes de corrections d'erreur. Enfin, les limites actuelles dans la synthèse d'ADN obligent les brins à ne pas dépasser 200 bases, ce qui nécessite des mécanismes supplémentaires pour permettre l'assemblage des données.
 - b. Le stockage par nanotechnologies ADN utilise quant à lui des structures d'ADN complexes pour stocker des informations. Relativement nouvelle, cette approche est moins mature et tous les enjeux n'en sont pas encore explorés. Cette méthode utilise des configurations comme l'origami d'ADN, où de courtes chaînes d'ADN sont pliées en structures spécifiques qui peuvent ensuite être utilisées pour représenter des données. Ces techniques offrent la possibilité de structurer les données de manière à en faciliter l'accès et la récupération sélective. Les structures d'origami d'ADN peuvent être conçues pour intégrer des motifs reconnaissables qui simplifient la lecture des données à l'aide de méthodes non séquentielles, telles que la microscopie.
3. Préservation : cette étape consiste à s'assurer de la conservation de l'information. Pour cette étape, il est crucial de prendre en compte la durée de conservation, le coût, spécialement énergétique, et l'impact du stockage sur la densité et l'accès à l'information.
 - a. Dans le cas du stockage par séquence, on peut identifier trois grandes familles de techniques : la déshydratation, l'encapsulation et la préservation in-vivo. Une matrice des différentes possibilités est présentée par Dorrichi et al. (2022, tab. 1).
 - b. Dans le cas du stockage par nanotechnologie, Wang et al. (2024, p. 10) rapportent des développements par lyophilisation ou cryogénéisation.

4. Accès direct : cette étape se réfère à la sélection des données. Dans n'importe quel système de stockage, il est crucial que la requête retourne la bonne donnée dans un temps limité. Contrairement à un système de fichiers traditionnel, une des limitations majeures du stockage sur ADN est l'absence d'adressage ou d'index.
 - a. Lorsque l'information est stockée sur des séquences de nucléotides, des techniques à base d'amplification PCR sur la base d'amorces placées de chaque côté du brin permettent d'accéder à n'importe quelle information. L'amplification par PCR présente toutefois des désavantages : la nécessité d'ajouter des amorces à chaque brin diminue la quantité d'information effective contenue par le brin ; de plus, chaque amplification fait disparaître un certain nombre de brins du réservoir (Erlich, Zielinski 2017, p. 3). D'autres méthodes basées sur de l'isolation physique ou du microcompartimentage permettent également d'accéder à n'importe quel document dans le réservoir, voire de faire des recherches sur la base des métadonnées (Wang et al. 2024, pp. 5-7).
 - b. Par stockage par nanotechnologie, Wang et al. (2024, p. 10) notent que l'accès direct est encore relativement inexploré, mais qu'il pourrait se faire par PCR.
5. Lecture : cette étape implique l'interprétation précise des structures ou séquences d'ADN dans lesquelles les données numériques ont été encodées. Il est vital d'éviter les erreurs de lecture sur un support et de s'assurer que tous les supports se rapportant à un même document soient lus
 - a. Dans le cas du stockage par séquences d'ADN, la lecture est réalisée par des techniques de séquençage d'ADN, qui permettent de déterminer l'ordre des nucléotides (Wang et al. 2024, pp. 5-7).
 - b. Dans le cas du stockage par nanotechnologies, cela nécessite des technologies telles que la microscopie ou des techniques spécifiques de visualisation moléculaire pour identifier les motifs structuraux uniques utilisés pour stocker les données.
6. Décodage : cette étape finale transforme les séquences de nucléotides ou les motifs structuraux en données numériques originales. Comme précisé précédemment, des algorithmes spécifiques sont nécessaires pour traduire correctement ces informations en présence d'erreurs potentielles introduites lors des étapes précédentes.

3. Le modèle OAIS

3.1 Historique et justification

Le modèle OAIS (Open Archival Information System) a été développé par le Consultative Committee for Space Data Systems (CCSDS), une organisation fondée en 1982 et rassemblant les agences spatiales nationales de différents pays collaborant pour le développement de standards de stockage de données pour la recherche spatiale. Confrontées à une masse de données numériques toujours croissante, ces agences étaient également soucieuses de les préserver à long terme. Dès 1990, le CCSDS conclut un accord avec l'Organisation Internationale de Normalisation (ISO) pour que ses recommandations suivent les procédures de normalisation de l'ISO (Lavoie 2014, p. 5).

Rapidement, le CCSDS est confronté à l'absence de tout consensus sur un modèle ou un cadre à utiliser pour la préservation à long terme des données numériques. Dès 1995, le CCSDS prend la décision de travailler à la création d'un modèle de référence qui viendrait définir les composants de base d'un système de préservation à long terme, les acteurs internes et externes de ce système ainsi que la nature des Objets-information qui seront gérés. Ces définitions sont exprimées à l'aide d'une terminologie spécifique à même de transcender tout vocabulaire spécifique propre à un domaine d'activité. En outre, le modèle de référence énumère les prérequis nécessaires au fonctionnement d'un tel système (Lavoie 2014, p. 5).

Dans sa conception, le modèle OAIS peut également s'appliquer à la préservation à long terme d'objets physiques. C'est toutefois pour son application aux documents numériques qu'il a été pensé et pour laquelle il est le plus connu et c'est dans cette optique que nous l'aborderons tout au long de ce mémoire. Il faut également noter que le modèle de référence a été dès le début pensé pour être applicable à tout type d'organisation, sans se limiter aux archives traditionnelles. Cette approche inclusive a permis l'édition d'un modèle reconnu par, et applicable à, tout type de secteur d'activité (Lavoie 2014, p. 5).

Les premières propositions ont été rédigées dès mai 1997 et le modèle est approuvé en tant que proposition de rédaction ISO en juin 2000. Après d'ultimes révisions, le modèle est approuvé en janvier 2002 en tant que norme ISO 14721 et officiellement publié en 2003. La norme publiée est strictement identique au Livre magenta (Magenta Book) émis par le CCSDS pour son usage interne. En accord avec les politiques tant de l'ISO que du CCSDS, la norme est périodiquement réévaluée. Un processus commencé en 2006 a abouti à la publication d'une version révisée de la norme ISO en 2012 (ISO 2012). B. Lavoie juge les révisions modestes, avec quelques exceptions (Lavoie 2014, p. 6).

Depuis la revue de B. Lavoie, la norme a continué d'évoluer. En date, la dernière proposition du CCSDS date de septembre 2019. Celle-ci inclut certaines additions et clarifications ; par rapport à ce travail, les plus intéressantes concernent la relation entre l'Information de préservation et l'Objet-information ainsi que la définition du Paquet d'informations (CCSDS 2019, p. v). La norme ISO est présentement soumise à réévaluation. Le modèle OAIS fait également l'objet de développement par la communauté professionnelle de la préservation numérique, qui a abouti par exemple au modèle Outer OAIS-Inner OAIS (Zierau 2017).

3.2 Concept

3.2.1 Principes généraux

L'acronyme du modèle OAIS peut se décomposer en deux parties. D'une part, le terme *Open* fait référence aux modalités de développement du modèle de référence, qui ont eu lieu à travers des forums ouverts où toute partie intéressée pouvait contribuer. Le modèle souligne que le terme ne présuppose aucunement du niveau d'accessibilité des documents archivés, qui peuvent être confidentiels et non accessibles au public (CCSDS 2019, sect. 1.1). D'autre part, l'*Archival Information System* se réfère à une organisation, composée de personnes et de systèmes, œuvrant à la préservation à long terme d'informations pour une communauté spécifique (CCSDS 2019, sect. 1.1). Comme souligné par B. Lavoie, cette définition met l'accent sur deux aspects : la préservation à long terme et l'accès aux documents archivés (Lavoie 2014, p. 7).

En tant que modèle de référence, le modèle OAIS se contente de définir un cadre d'application et se veut agnostique vis-à-vis de toute architecture ou d'implémentation spécifique, allant jusqu'à suggérer qu'une implémentation pourrait peut modifier l'agencement de certaines fonctions en fonction de ses besoins ou de ses moyens.

Ci-après, lorsque le terme « OAIS » sera employé seul, celui-ci fera référence à l'archive elle-même. Le terme « modèle OAIS » fait référence au modèle tel qu'il est décrit dans les différents documents du CCSDS. Les termes utilisés par le modèle OAIS suivront la traduction française officielle¹ et seront systématiquement écrits avec une majuscule ; leur premier emploi sera suivi par le terme original anglais en italiques et entre parenthèses.

3.2.2 Environnement

Une OAIS ne saurait exister pour elle-même. Elle se place à la croisée de trois groupes d'acteurs (CCSDS 2019, sect. 2.1) :

1. Les Producteurs (*Producers*) sont ceux qui transmettent à l'archive les documents à conserver.
2. Les Utilisateurs (*Consumers*) sont ceux qui interagissent avec l'archive pour consulter des documents d'intérêt. Les Utilisateurs peuvent également être des Producteurs. Un groupe particulier d'Utilisateurs est la Communauté d'utilisateurs cible (*Designated Community*), pour laquelle l'OAIS est adaptée. En particulier, ce groupe d'utilisateurs est en mesure de comprendre directement les informations sauvegardées par l'OAIS.
3. Le Management (*Management*) est responsable de la mise en place d'une archive et des politiques s'y afférant. Il n'est pas responsable de la gestion courante de l'OAIS.

3.2.3 Entités fonctionnelles

Au sein de l'OAIS, six Entités fonctionnelles (*Functional entities*) liées entre elles se partagent différentes responsabilités (CCSDS 2019, sect. 4.1).

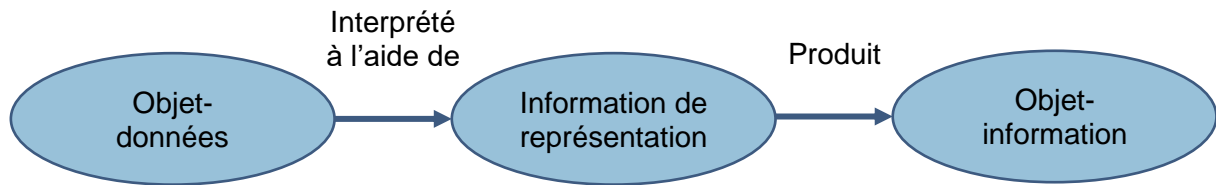
¹ Cette traduction, datée d'octobre 2017, est disponible à l'adresse [https://public.ccsds.org/Pubs/650x0m2\(F\).pdf](https://public.ccsds.org/Pubs/650x0m2(F).pdf). Elle est basée sur la version 2012 du modèle OAIS et ne comprend donc pas les modifications ultérieures de 2019. Elle ne sera utilisée dans le présent travail que pour la traduction des termes. Toutes les références au modèle OAIS se font à travers l'original anglais.

1. L'entité Entrée (*Ingest*) est responsable de l'acceptation de l'information du Producteur et de sa préparation pour le stockage et la gestion au sein de l'archive. Cela comprend des tâches telles que la validation des données, la normalisation et l'attribution d'identifiants uniques.
2. L'entité Stockage (*Archival storage*) est l'endroit où les informations reçues sont stockées et conservées au fil du temps. Cette entité fonctionnelle assure la préservation à long terme des ressources numériques, y compris leur intégrité et leur authenticité. Elle gère les ressources de stockage et fournit des mécanismes de récupération et d'accès en cas de besoin.
3. L'entité Gestion des données (*Data management*) supervise l'organisation, la description et le contrôle des informations archivées. Elle maintient les métadonnées relatives aux objets stockés, permettant la recherche, la sélection et l'accès au contenu. Cette entité s'occupe également de la gestion des versions, de la migration et d'autres tâches de gestion du cycle de vie des données.
4. L'entité Administration (*Administration*) englobe la gestion globale et la coordination des activités de l'archive. Elle comprend des fonctions telles que l'élaboration des politiques, l'allocation des ressources, le contrôle des performances du système et la garantie du respect des normes et des réglementations en vigueur. Cette entité agit en tant que couche de gouvernance pour l'ensemble du système d'archivage.
5. L'entité Accès (*Access*) facilite l'interaction de l'utilisateur avec l'information archivée. Elle fournit des mécanismes de recherche, de navigation et d'extraction du contenu en fonction des requêtes des Utilisateurs. Les contrôles d'accès et les mécanismes d'authentification garantissent que les Utilisateurs ne peuvent accéder qu'au contenu autorisé conformément aux politiques et aux permissions définies.
6. L'entité Planification de la pérennisation (*Preservation planning*) implique le développement de stratégies et de procédures pour assurer la viabilité à long terme de l'information archivée. Cette entité évalue les risques pour la préservation numérique, identifie les exigences de préservation et définit des stratégies et des actions de préservation pour atténuer ces risques au fil du temps. Elle vise à maintenir l'intégrité, l'authenticité et la facilité d'utilisation des objets archivés pour les générations futures.

3.2.4 Information

Nous avons jusqu'ici utilisé le terme « information » de manière générale et sans le définir explicitement. Le modèle OAIS définit l'information comme tout type de connaissance pouvant être échangée ; à cet effet, l'information est exprimée à travers une forme ou une autre, appelée donnée. Pour pouvoir être intelligible et transmettre l'information qu'elle exprime, la donnée doit être interprétée à travers une certaine Base de connaissance (*Knowledge Base*), propre à une personne ou à un ensemble de personne. Si cette Base de connaissance fait défaut, alors la donnée doit être accompagnée d'une Information de représentation (*Representation Information*) permettant à qui la consulte d'interpréter la donnée pour comprendre l'information transmise (CCSDS 2019, sect. 2.2.1). Dans le cas d'un Objet numérique (*Digital Object*), il s'agit donc de convertir la séquence de bits en information porteuse de sens. On peut résumer cela avec le schéma suivant :

Figure 2. Lien entre information et données dans le modèle OAIS



(adapté de CCSDS 2019, fig. 2-2)

Les questions d'interprétation auxquels répond l'Information de représentation peuvent être multiples (CCSDS 2019, sect. 4.2.1.3). Globalement, on peut distinguer trois types. Le premier, appelé Information de structure (*Structure Information*), permet d'identifier comment l'Objet-données doit être interprété : il décrit ainsi si la séquence de bits forme des nombres ou des lettres, des pixels ou des tableaux, etc. De manière plus complexe, l'Information de structure détermine le format de l'Objet-numérique. L'Information de structure étant rarement suffisante par elle-même, elle est souvent associée à un deuxième type, appelée Information sémantique (*Semantic Information*). Celle-ci va venir ajouter des précisions sur le sens des données, par exemple la langue d'une chaîne de caractères ou la nature des nombres, et sur leurs liens éventuels. Le modèle OAIS souligne que l'Information sémantique est indépendante du format utilisé par l'Objet numérique. Enfin, le troisième type est défini comme une Autre information de représentation (*Other Representation Information*) et comprend toute Information de représentation dont la classification échappe aux deux autres : le modèle OAIS prend notamment pour exemple un logiciel permettant de lire le format utilisé par l'Objet numérique.

Le modèle OAIS précise que pour être préservée à long terme, l'information doit être maîtrisée au niveau de l'Objet-données et de l'Information de représentation nécessaire à son interprétation. Or, l'Information de représentation, particulièrement quand elle est numérique, est elle-même constituée de ses propres données et de sa propre Information de représentation. Il se crée donc un Réseau de représentation (*Representation Information Network*), une chaîne récursive d'Information de représentation jusqu'à arriver à une Information de représentation exprimée par une forme physique interprétable directement par la Communauté d'utilisateurs cible (CCSDS 2019, sect. 4.2.1.3.2). C'est la Base de connaissance de cette dernière qui détermine le niveau d'Information de représentation minimum nécessaire au fonctionnement courant de l'OAIS, bien que celle-ci soit libre d'aller au-delà et de maintenir une Information de représentation plus étendue afin de toucher un nombre plus large d'Utilisateurs. Dans certains cas, la Base de connaissance d'une Communauté d'utilisateurs cible spécifique est suffisante en tant que telle pour se passer d'Information de représentation. Toutefois, le modèle OAIS suggère qu'il est préférable d'acquérir toute Information de représentation qui peut l'être lors de l'entrée d'un document dans l'OAIS, car il sera probablement plus coûteux de la redécouvrir ultérieurement en cas de changement dans la Base de connaissance (CCSDS 2019, sect. 4.2.1.3). Il laisse toutefois ouverte la possibilité qu'une OAIS se contente de référencer l'existence dans une autre OAIS de toute Information de représentation nécessaire (CCSDS 2019, sect. 2.2.1).

Pris ensemble, l'Objet-données et l'Information de représentation forment l'Information de contenu (*Content information*). L'information de contenu représente ce qui doit être préservé par une OAIS. Le modèle OAIS souligne que décider ce qui est du ressort de l'Objet-données ou de l'Information de représentation n'est pas nécessairement évident et peut être adapté à

la discrétion de chaque OAIS (CCSDS 2019, sect. 4.2.1.4.1). Crucialement, ce choix peut être fait dès l'entrée des documents et l'OAIS peut décider de simplifier l'Objet-données pour économiser des étapes dans le Réseau de représentation, par exemple en extrayant le texte d'un document pour ne pas conserver un format particulier.

L'Information de contenu ne circule jamais seule au sein d'une OAIS. En effet, elle est systématiquement associée à d'autres éléments pour former un Paquet d'informations (*Information Package*). Celui-ci contient, outre l'Information de contenu, une Information de pérennisation (*Preservation Description Information (PDI)*) qui permet d'identifier l'Information de contenu et l'environnement dans lequel elle a été créée (CCSDS 2019, sect. 4.2.1.4.2). Il est possible de comprendre l'Information de pérennisation comme des métadonnées sur l'Information de contenu et en particulier sur l'Objet-données. En ce sens, sa présence dans le paquet est obligatoire, bien que la portée de son contenu soit laissée à la discrétion de l'OAIS. L'Information de pérennisation comprend :

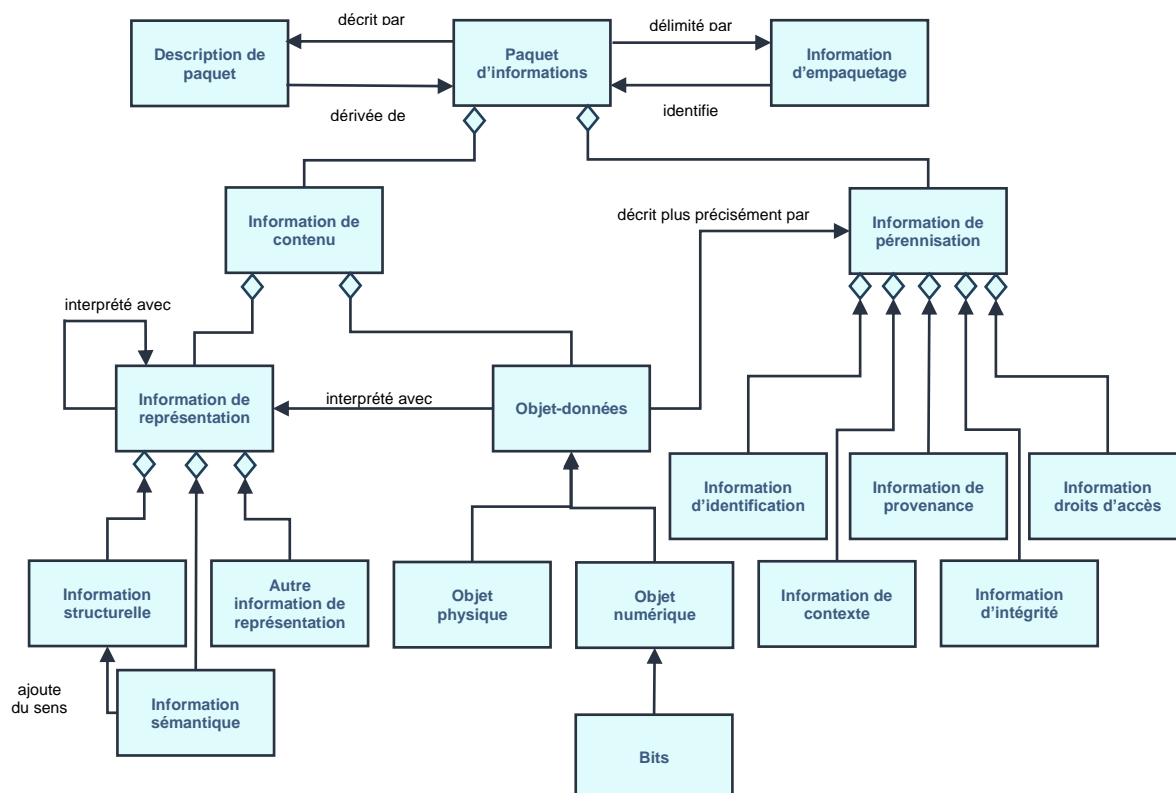
- l'Identification (*Reference Information*), qui permet de faire référence de manière non ambiguë au contenu ;
- le Contexte (*Context Information*), qui décrit les relations entre l'Information de contenu et l'environnement où elle a été créée ;
- la Provenance (*Provenance Information*), qui documente l'histoire de l'Information de contenu, de sa source à ses changements éventuels ;
- l'Intégrité (*Fixity Information*), qui offre des mécanismes de vérification que l'Information de contenu est bien celle qu'elle prétend être et qu'elle n'a pas été altérée ;
- les Droits d'accès (*Access Rights Information*), qui identifient les restrictions à la consultation applicables à l'Information de contenu (CCSDS 2019, sect. 2.2.2).

Le paquet lui-même est associé à une Information d'emballage (*Packaging Information*), qui décrit comment les différents éléments du paquet sont liés entre eux, comment les identifier et comment les extraire ; il peut également avoir sa propre Information de description du paquet (*Package Description Information*) (CCSDS 2019, sect. 4.2.1.4.3).

Cette Information d'emballage ne doit pas être confondu avec la Description de paquet (*Package Description*), qui décrit le contenu du Paquet. La Description de paquet alimente les Outils d'accès (*Access Aids*) de l'OAIS, permettant à l'Utilisateur de localiser, d'identifier et de sélectionner l'information pertinente à son usage (CCSDS 2019, sect. 4.2.1.4.4). En tant que telle, la Description de paquet n'est pas nécessaire à la préservation à long terme et son contenu est dépendant de l'Information de contenu et de l'Information de pérennisation (CCSDS 2019, sect. 4.2.2.3).

L'Information de contenu, l'Information de préservation, l'Information d'emballage et la Description du paquet sont des types spécialisés d'Objet-information (*Information object*).

Figure 3. Schéma des différents composants du Paquet d'informations



(adapté de CCSDS 2019, fig. 4-19)

Un paquet conservé par une OAIS est dénommé Paquet d'informations archivé (*Archival Information Package (AIP)*). Celui-ci peut être sensiblement différent du paquet transmis par le producteur, dénommé Paquet d'informations à verser (*Submission Information Package (SIP)*), que ce soit parce que celui-ci dispose d'Information de représentation ou d'Information de préservation insuffisantes ou que l'OAIS organise différemment les paquets. Comme précisé par le modèle OAIS, la relation entre un SIP et un AIP n'est pas nécessairement d'un pour un. En effet, un SIP peut être partagé en plusieurs AIPs, ou un AIP recomposé à partir de plusieurs SIPs. Enfin, un Utilisateur demandant un AIP se voit remettre un Paquet d'informations à verser (*Dissemination Information Package (DIP)*), qui peut ne pas contenir toutes les informations présentes au sein de l'AIP (CCSDS 2019, sect. 4.2.2.2). Ces paquets diffèrent dans les contenus obligatoires qu'ils doivent comporter ainsi que dans les associations entre ces contenus. Dans le cas de ce mémoire, nous ne nous intéresserons qu'aux AIPs.

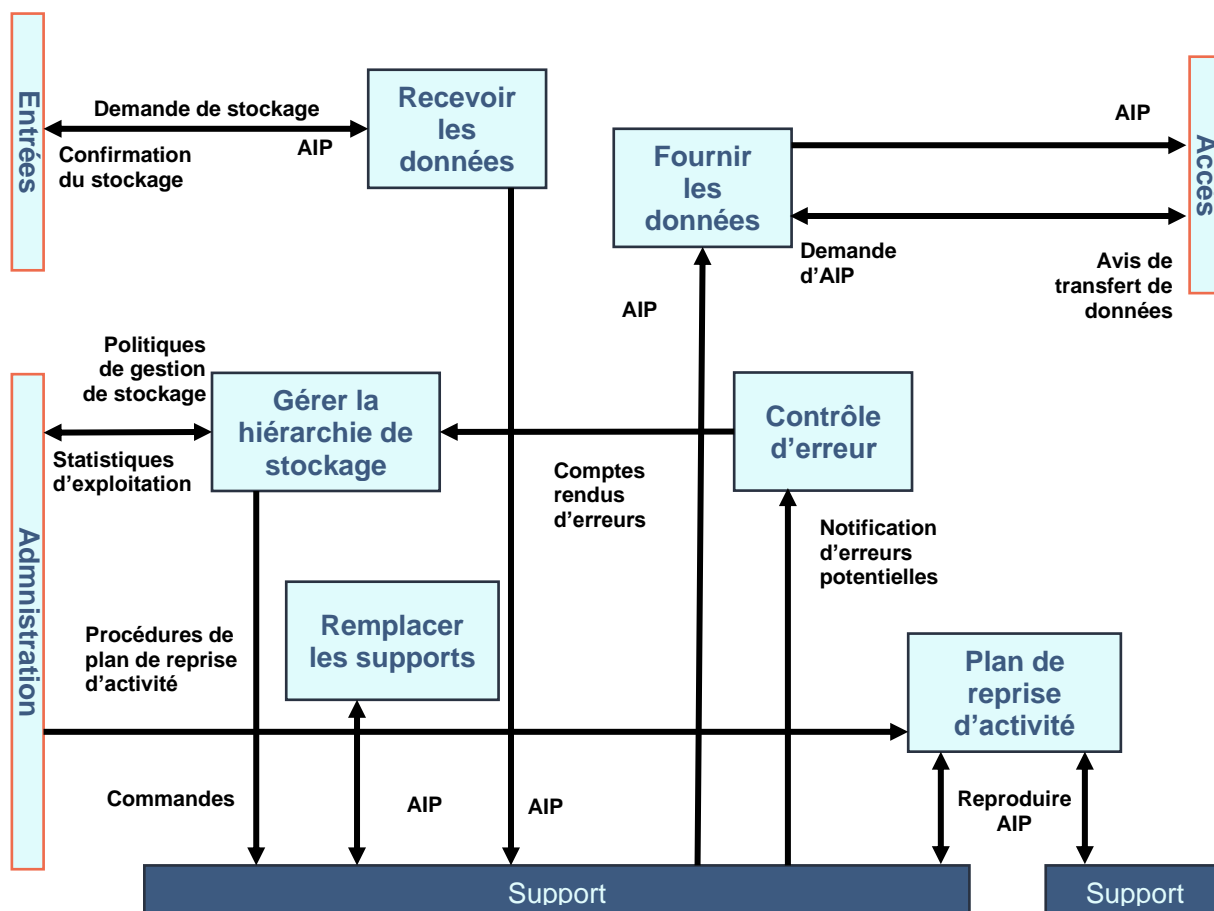
Pour être découvrable au sein de l'OAIS, un AIP doit être référencé par un ou plusieurs Outils d'accès. L'information nécessaire pour ces derniers est appelée Description associée (*Associated Description*), qui forme également la base des Descriptions de paquet dont nous avons parlé plus haut, chaque Description de paquet étant composée d'une ou de plusieurs Descriptions associées. Les Outils d'accès utilisent l'Identification fournie par l'Information de pérennisation pour identifier l'AIP.

3.3 Entité fonctionnelle Stockage

Dans le cadre de ce mémoire, nous nous intéresserons particulièrement à l'entité fonctionnelle « Stockage » (ISO 2012, sect. 4.1.1.3), puisque c'est au sein de celle-ci que l'ADN peut être

amenée à jouer un rôle. Située au cœur de l'OAIS, l'entité fonctionnelle Stockage ne gère que des AIPs. Elle est en contact avec les entités fonctionnelles Entrées, Accès et Administration.

Figure 4. Schéma de l'entité fonctionnelle Stockage



(adapté de CCSDS 2019, fig. 4-3)

Ce schéma présente les six fonctions de l'entité Stockage et leurs relations éventuelles avec les autres entités, de même qu'avec le ou les mécanismes de stockage d'informations numériques, locaux ou distants, indistinctement rassemblés sous le terme de « Support ». Dans le cas du présent mémoire, l'ADN représentera un de ces supports.

La fonction *Recevoir les données* (*Receive data*) reçoit de l'entité Entrées une demande de stockage (*Storage request*) avec un AIP. Cette fonction sélectionne le support de stockage permanent le plus adapté en fonction des caractéristiques de l'AIP, y compris sa fréquence d'utilisation prévue, et l'y transfère. Une fois le transfert terminé, cette fonction envoie un message de confirmation de stockage (*Storage confirmation*) à l'entité Entrées.

La fonction *Gérer la hiérarchie de stockage* (*Manage Storage Hierarchy*) distribue par des Commandes (*Commands*) les AIPs sur les supports adaptés, selon les Politiques de gestion de stockage (*Storage management policies*) et les Statistiques d'exploitation (*Operational statistics*). Cette fonction s'adapte également à toute particularité d'un AIP vis-à-vis de son stockage (par exemple la sécurité). En outre, elle surveille les Comptes-rendus d'erreur (*Error logs*) pour s'assurer que les AIPs ne sont pas corrompus. Enfin, elle fournit des Statistiques d'exploitation à l'entité Administration.

La fonction Remplacer les supports (*Replace media*) s'assure que les AIPs peuvent être reproduit au cours du temps. Sans altérer ni l'Information de contenu ni l'Information de pérennisation, cette fonction peut modifier l'Information d'emballage pour que celle-ci puisse continuer sa fonction sur des supports différents.

La fonction Contrôle d'erreur (*Error checking*) fournit une assurance statistiquement acceptable qu'aucun composant de l'AIP ne s'est corrompu dans le support ou lors d'un transfert de données interne à l'OAIS. Cette fonction nécessite que tous les composants matériels ou logiciels fournissent des rapports d'erreurs qui puisse être redirigés vers un système standardisé qui soit régulièrement vérifié. L'Information de préservation fourni des assurances que le contenu n'a pas été altéré lors du mouvement ou de l'accès, mais des informations similaires sont nécessaires pour protéger l'Information de préservation elle-même.

La fonction Plan de reprise d'activité (*Disaster recovery*) fournit un mécanisme permettant de dupliquer le contenu numérique de la collection d'archives permettant, le cas échéant, de reconstituer les fonds après un sinistre. Les détails des Procédures de plan de reprise d'activité (*Disaster recovery policy*) sont spécifiés par l'entité Administration.

La fonction Fournir les données (*Provide Data*) fournit des copies des AIPs à l'entité d'Accès lorsqu'elle est sollicité par une Demande d'AIP (*AIP request*). La fonction renvoie également un Avis de transfert de données (*Notice of data transfert*) à l'issue d'une demande.

4. Revue de la littérature sur la préservation à long terme

4.1 Introduction

Le chapitre précédent, décrivant certains concepts du modèle OAIS, a permis d'esquisser le cadre théorique dans lequel se place la réflexion de ce mémoire. Le but du présent chapitre est de recenser les mécanismes concrets qui ont pu être mis en œuvre pour la préservation à long terme des données numériques, indépendamment de leur référence ou non à ce modèle.

La préservation à long terme, que ce soit sur des années, des décennies, des siècles ou au-delà, a pour objectif qu'entre l'archivage des documents et le jour où un utilisateur voudra y accéder à nouveau, celui-ci sera capable d'accéder aux documents et d'interpréter correctement l'information qu'ils contiennent. Naturellement, selon les échelles de temps considérées, les enjeux de la préservation à long terme ne sont pas tout à fait les mêmes. Dans le cadre de cette revue de la littérature, nous considérerons le temps long selon la définition du modèle OAIS, à savoir :

« Long Term is long enough to be concerned with the impacts of changing technologies, as well as support for new media and data formats, or with a changing Knowledge Base of the Designated Community or changes within the Designated Community or its definition. Long Term may extend indefinitely. » (CCSDS 2019, sect. 1.1)

Dans le cadre de l'archivage sur ADN, et partant du cas de figure où tout doit être redécouvert, la préservation à long terme peut se structurer en trois étapes : premièrement, la préservation à long terme de la partie immédiatement perceptible de l'archive, à savoir sa documentation, ses libellés et plus largement sa nature même d'archive, le tout permettant d'atteindre les nucléotides en tant que support d'information ; deuxièmement, la préservation à long terme du support permet, depuis les nucléotides, d'atteindre les bits ; troisièmement, la préservation à long terme des données permet d'atteindre, depuis les bits, l'information. Dans le cadre du stockage sur ADN, la seconde étape peut ne pas déboucher sur des bits, mais directement sur l'information elle-même si le codec utilisé transcrit des bits. Dans les sections suivantes, nous aborderons successivement ces trois étapes.

4.2 Préservation de la partie immédiatement perceptible

4.2.1 Introduction

Contrairement à un support analogique, le contenu d'un support de stockage numérique est entièrement inconnu jusqu'à ce qu'il soit lu. Alors qu'il peut aller de soi que des rangées d'étagères remplies de documents constituent une archive, il n'est en réalité pas si évident qu'un CD-ROM est un support de stockage (Rothenberg 1998, pp. 1-2) : on en voit plus souvent aujourd'hui effrayer les oiseaux dans un potager que transmettre une information. Il en va de même pour un tube à essai dont un utilisateur futur pourrait ne pas savoir s'il contient un dangereux bacille ou des documents. Mais une fois le caractère informationnel de l'objet reconnu, il se pose encore la question de la manière d'y accéder : dans l'exemple de J. Rothenberg, c'est une lettre manuscrite en anglais qui sert de médiateur pour accéder à l'objet numérique. Une transmission adéquate de contenu numérique ne saurait donc se passer entièrement de l'usage d'un support analogique interprétable sans technologie particulière (Rothenberg 1999, p. 29; CCSDS 2019, sect. 4.2.1.3.2).

Cette section s'intéresse donc à la partie immédiatement perceptible de l'archive, c'est-à-dire tout ce qu'un être humain peut immédiatement comprendre sans intermédiaire technologique

et qui, *in fine*, lui permettra d'accéder au contenu qui n'est pas immédiatement perceptible. Après une brève section sur les libellés, le cœur de cette partie passe en revue les différentes propositions qui ont été faites sur les questions liées à la documentation, en nous intéressant tant aux questions du support de cette documentation que des informations qu'elle doit inclure.

4.2.2 Libellés

Il n'est pas impossible que le premier contact avec une archive se fasse par l'intermédiaire de courtes suites de lettres ou de symboles. Si à notre connaissance aucun travail n'a été mené sur la sémiologie utilisable par une archive, plusieurs études ont abordé la question des libellés (Rothenberg 1999, pp. 29-30; LaBarca 2012, p. 137). Un travail plus précisément orienté sur ces questions étant actuellement réalisé par S. Valot, nous ne poussons pas plus loin nos réflexions et renvoyons les lecteurs à son travail.

4.2.3 Documentation

4.2.3.1 Principes généraux

4.2.3.1.1 Choix du support analogique

La nécessité de transmettre des informations via des supports analogiques pose évidemment la question du type de support. Celui-ci doit être suffisamment résistant pour permettre sa conservation à côté d'un support de stockage sur ADN. En outre, il devrait autant que possible se passer de technologies trop particulières et permettre une certaine densité d'information.

L'Agence nationale pour la gestion des déchets radioactifs (ANDRA) a recours au papier permanent (ISO 1994) pour le stockage à long terme des documents (Dumont, Charton 2012) et c'est également la solution retenue par la Nuclear Energy Agency (Nuclear Energy Agency 2019a, p. 38), tandis que la solution Micr'Olonys développée par la société Eupalia prévoit l'usage soit du papier permanent, soit du microfilm (Appuswamy, Joguín 2021) ; de son côté, l'entreprise Piql a développé un film polyester (Sabliński, Trujillo 2021, pp. 23-25). Ces trois supports offrent des durées de vie allant de plusieurs centaines d'années à un millénaire, sauf dégradation par processus chimiques, et permettent un stockage modérément dense d'information qu'il est possible d'augmenter en procédant à la miniaturisation du contenu.

Alternativement, des supports solides, minéraux ou métalliques, plus résistants aux accidents et offrant une durée de vie de plusieurs milliers d'années, ont également été utilisés. C'est par exemple le cas des plaques emportées par les sondes Pioneer 10 et 11 et Voyager 1 et 2 (Capova 2021) ; le choix du métal pour ces plaques est dû aux phénomènes d'érosion par impacts de micrométéorites (Sagan, Sagan, Drake 1972, p. 881). Toutefois, la densité d'information de tels supports, difficilement superposables, est très faible et la miniaturisation du contenu est nécessaire pour pouvoir stocker une quantité importante d'information. Ainsi, la fondation Long Now a fait développer par le Laboratoire national de Los Alamos et la société Norsam Technologies des disques de nickel (Kelly 2008); l'ANDRA a quant à elle eut recours à des micro-disques de saphir (Dumont, Charton 2012; Rey 2013). Dans ces deux cas, il est prévu que l'objet dure des milliers d'années. En outre, l'information y est stockée sous forme d'image représentant des pages de texte : un agrandisseur optique et une connaissance de la langue est tout ce qui est nécessaire pour accéder à l'information. De manière ingénieuse, une des faces du disque de la fondation Long Now suggère la miniaturisation par l'inclusion de texte de plus en plus petit.

Toutefois, dans le cas des disques de saphir, plusieurs se sont brisés lors de leur manipulation (Calla et al. 2023, p. 5). Il paraît également opportun de souligner que de nombreux objets en métal fabriqués par les sociétés passées ont été refondus pour en récupérer la matière première : ainsi, la solidité annoncée de tels supports ne prend pas en compte les dégâts humains. Enfin, il faut également souligner le coût plus élevé de tels supports par rapport au papier ou au microfilm.

4.2.3.1.2 *Mode de communication et langage*

La seule conservation du support d'information ne garantit pas l'intelligibilité des données qu'il transmet et plusieurs cas dans l'histoire de l'humanité attestent de cet état de fait, par exemple le linéaire A en Crète, indéchiffré à ce jour. Le choix du langage est donc crucial et il n'est pas étonnant qu'une des premières questions auxquelles a été confrontée l'ANDRA en créant ses micro-disques de saphirs, au-delà des aspects d'ingénierie, a été la langue à utiliser pour transmettre des informations (Dumont, Charton 2012). Si dans bon nombre de cas de figures l'absence de toute information à ce propos semble suggérer le choix d'une seule langue, quand la question est abordée, plusieurs projets font le choix d'écrire la documentation en plusieurs langues (Nuclear Energy Agency 2019b, pp. 19-21; Joguín, Dumont 2022).

Dans la perspective des communications interstellaires, que l'on peut extrapoler comme des communications intergénérationnelles (Jiang et al. 2023, pp. 3, 6), il n'est pas question d'utiliser un quelconque langage d'origine humaine : toutes les tentatives de communications avec d'autres formes de vies se sont basées sur des éléments auditifs, graphiques, mathématiques ou physiques (Sagan, Sagan, Drake 1972; Capova 2021; Jiang et al. 2023, p. 2). Le choix de proposer une variante de la documentation reposant uniquement sur les mathématiques peut également être relevé dans le projet de machine abstraite porté par Rummelhoff et al. (2021). Le projet Message In A Bottle (MIAB) se base sur une « représentation symbolique » pour transmettre des instructions à des intelligences extraterrestres, mais ne détaille pas en quoi elle consiste (Jiang et al. 2023, p. 6).

En ce qui concerne les langages informatiques, plusieurs études proposent d'utiliser du pseudo-code afin de ne pas dépendre d'un langage en particulier dont la pérennité n'est pas assurée sur le long terme (Nguyen, Kay 2015; Appuswamy, Joguín 2021, pp. 5-6; Andriamahady 2021, p. 48; Joguín, Dumont 2022, p. 273).

4.2.3.1.3 *Technicité*

Les questions de langue traitées, il reste à aborder le niveau de langage et de manière plus générale la technicité de la documentation.

La Nuclear Energy Agency prévoit ainsi des niveaux de technicités différents selon le type de document, privilégiant un langage simple pour les documents de plus haut niveau et une technicité accrue pour des documents plus spécifiques (Nuclear Energy Agency 2019a, pp. 64-66).

Le projet de machine virtuelle mené par I. Rummelhoff et al. fait quant à lui un choix différent en proposant trois versions de la même documentation mais écrites dans un niveau de technicité différent, l'un en langage simple, l'autre plus complexe et le troisième purement mathématique (Rummelhoff et al. 2021, pp. 154915-154916). N. T. Nguyen et A. Kay ont quant à eux prévu que les instructions pour leur machine virtuelle soient le plus réduites possible et relèvent qu'une feuille A4 a une « qualité psychologique » en termes de compacité et

d'élégance ; le but étant d'implémenter leur machine virtuelle en une « fun afternoon hack » (Nguyen, Kay 2015, p. 6).

4.2.3.2 Choix des documents à conserver

La préservation à long terme de l'archive en tant que vecteur de mémoire doit s'inclure dans une approche relativement holistique, prenant également en compte le contexte culturel et sociétal de l'archive. À notre connaissance, peu de secteurs d'activité ont poussé la réflexion aussi loin que celui de l'énergie atomique et en particulier celui de la gestion des déchets radioactifs, dont la mémoire nécessite d'être préservée sur des millénaires (Popham, Mitcham 2022, p. 315).

Le *Radioactive Waste Management Committee (RWMC)* de la *Nuclear Energy Agency (NEA)* est ainsi à l'origine d'un projet intitulé « *Preservation of Records, Knowledge and Memory Across Generations* » (*RK&M initiative*). Cette initiative avait pour but d'une part de développer une compréhension théorique générale de la préservation à long terme et d'autre part de concevoir une « boîte à outils » d'approches génériques mais concrètes pour y répondre (Nuclear Energy Agency 2019a, pp. 18-19). Cette initiative s'est conclue en 2019 par la publication d'un rapport faisant la synthèse des discussions ayant eu lieu dans ce secteur d'activité depuis 1972. Bien que le public cible du rapport reste la communauté professionnelle en charge de la gestion des déchets radioactifs, plusieurs concepts pertinents pour notre problématique y sont développés², mais parmi les trente-cinq mécanismes pour la préservation à long terme identifiés, peu sont directement transposables à la préservation numérique à long terme. Ces mécanismes sont répartis en neuf familles :

1. Ensembles de documents dédiés et documents récapitulatifs (*Dedicated record sets and summary files*)
2. Institutions de mémoires (*Memory institutions*)
3. Marqueurs (*Markers*)
4. Capsules temporelles (*Time capsules*)
5. Culture, éducation et art (*Culture, education and art*)
6. Gestion de la connaissance (*Knowledge management*)
7. Dispositifs de surveillance (*Oversight provisions*)
8. Mécanismes internationaux (*International mechanisms*)
9. Cadre réglementaire (*Regulatory framework*)

L'approche la plus pertinente proposée par le rapport nous semble être le premier point, à savoir la création de deux outils documentaires, le Fichier d'informations clés (*Key Information File (KIF)*) et l'Ensemble de documents essentiels (*Set of Essential Records (SER)*) (Nuclear Energy Agency 2019a, pp. 64-66; voir aussi respectivement 2019b; 2019c). L'application des autres approches mémorielles développées par la RK&M Initiative dépasse le cadre de ce mémoire, aussi ne les commentons-nous pas.

Le KIF est un document unique, produit de manière multidisciplinaire et participative, destiné à informer toute personne actuelle ou future sans nécessiter de connaissances spécialisées. Il doit être écrit de manière succincte et non-technique, hormis pour empêcher des

² Ci-après, nous considérerons les concepts développés du point de vue d'une archive ; il va de soi que toutes les discussions et tous les exemples du rapport ont pour un objet un site de gestion des déchets radioactifs.

interprétations ambiguës. Il transmet donc de manière accessible un aperçu de l'archive, de ses prises de décision et de ses intentions, mais devrait également pointer vers d'autres documents plus spécialisés, dont le SER. Il est précisé que le KIF doit être produit dans un format standardisé et en plusieurs langues, dont la langue locale et toute autre langue nationale ; des parties peuvent également faire l'objet d'une traduction dans d'autres langues, spécialement celles utilisées dans des instances internationales comme l'Organisation des Nations Unies. Sa structure interne est composée des éléments suivants : contexte, localisation, design, inventaire, sécurité, résumé, accès aux éléments détaillés. D'autres spécificités liées à sa publicité et sa distribution dans d'autres institutions ne nous concerneront pas ici, si ce n'est pour spécifier que plusieurs copies doivent en être faites (Nuclear Energy Agency 2019b, pp. 19-21).

Le SER est quant à lui un ensemble de documents uniques sélectionnés pendant la durée de vie de l'archive, visant à fournir des informations suffisantes pour en garantir une compréhension adéquate. Là où le KIF visait une audience non spécialisée, le SER cible des spécialistes. La constitution du SER doit être équilibrée entre la nécessité d'en limiter la taille pour des raisons de clarté et la transmission d'autant d'informations que nécessaire. À cet effet, la Nuclear Energy Agency propose la création d'une classification évaluant chaque document selon sa pertinence vis-à-vis des besoins futurs et des efforts nécessaires à la recréation des données qu'il contient.

Tableau 1. Classification proposée par la NEA pour constituer le SER

Relevance/effort		a) Some effort	b) Extremely high effort
Not relevant	0		
Nice to have	1		
Should have	2		SER
Must have	3	SER	SER

(adapté de Nuclear Energy Agency 2019c, tab. 3.1)

Il convient de relever que les documents composants le SER semblent n'être que ceux produits par l'archive elle-même ; en tant que tel, il est possible que des documents importants nécessaires à l'interprétation future de l'archive échappent à cette classification car provenant de tiers externes.

C'est dans la préservation de documents tels que le KIF et le SER, et plus largement de l'archive en général, que le recours aux capsules temporelles (Nuclear Energy Agency 2019a, pp. 71-73) nous semble le plus approprié. Les capsules temporelles sont des dispositifs de stockage conçus pour préserver des informations pendant une période de temps donnée ou indéfinie. Elles peuvent contenir des documents, des objets ou des dispositifs de stockage numérique et sont enterrées ou scellées dans des environnements contrôlés. Plus crucialement peut-être, elles pourraient permettre d'encapsuler l'archive et toute documentation nécessaire en un seul élément.

Enfin, il convient de relever que le rapport de la Nuclear Energy Agency insiste sur l'importance de la redondance pour la préservation de la mémoire à long terme : cette redondance est atteinte par la multiplication des lieux de stockage des documents, par les liens entre les documents et par des modes de communication multiples et différents (Nuclear Energy Agency 2019a, pp. 22-23).

En lien étroit avec le sujet du présent travail, le mémoire de bachelor de D. Andriamahady s'est intéressé à la relation entre ADN et OAIS (Andriamahady 2021). Ce mémoire se conclut par une série de recommandations, dont les différents documents et informations devant être conservés :

- Documentation générale sur les procédures du projet
- Structure des identificateurs uniques de l'AIP
- Procédure de conception des amorces (logiciel, température de fusion, courbe standard)
- Procédure de codage des AIPs en brins et procédure réciproque
- Code, bibliothèques :
 - Hash (sha256)
 - Bibliothèque : Reed-Solomon, ZIP
- Paramètres d'étalonnage utilisés
 - Dictionnaire bit vers ADN
 - Données aléatoires pour le masque
 - Paramètres des algorithmes Reed-Solomon entrelacés (2 jeux de paramètres)
 - Longueur des données pour chaque brin
 - Longueur de l'amorce
 - Longueur de l'espace pour stocker les identificateurs
- Protocoles de laboratoire :
 - Processus d'accès à l'information
 - Procédure de lecture de l'ADN

4.3 Préservation du support

4.3.1 Introduction

Les problématiques liées à la préservation du support dans un environnement où ils deviennent obsolètes en quelques années sont déjà anciennes. Sans prétendre avoir fait le tour de la littérature, on peut déjà en trouver trace dès les années 1980. Ainsi, J. Mallinson (1986), résumant les délibérations du *Subcommittee C of the Committee on Preservation of the National Archives and Records Administration* des États-Unis, était déjà conscient de la futilité de développer des supports avec une longue durée de vie si le matériel permettant de les lire n'était plus accessible. Considérant l'absence d'interopérabilité entre les différentes machines et les difficultés financières et techniques liées à leur réparation, le *Subcommittee C* concluait que la solution à l'obsolescence du support était la migration en masse sur microfilm comme support résistant au temps et non dépendant d'une machine pour être lu, si ce n'est d'un agrandisseur optique. Une telle approche était résolument écartée par S. Gavrel (1986) qui, en réponse à cet article, notait déjà que la migration d'un support numérique à un support

analogique entraînait la perte irréversible de fonctionnalités et de la lisibilité du document par des machines, tout en reconnaissant que la migration entre supports était un processus minutieux et coûteux. Le même avis contre l'impression est exprimé par J. Rothenberg et D. Bearman (Rothenberg 1999, pp. 3-4; Bearman 1999).

Aujourd'hui, les 175 zettabytes annoncés pour 2025 (Reinsel, Gantz, Rydning 2018) et les problèmes de densité de stockage qu'ils posent déjà sous forme numérique rendent leur stockage sur microfilm irréaliste, sans parler des comportements interactifs ou multimédia inhérents au numérique (Heminger, Kelley 2005, pp. 13-14). Malgré des propositions plus récentes en ce sens (par exemple LaBarca 2012, p. 138), cela semble être une piste qui doit être définitivement écartée pour n'importe quelle archive d'importance. Cette quantité rend également peu souhaitable la migration périodique des supports en raison de leur vitesse d'évolution : sans un engagement constant, une carence dans la chaîne de migration peut devenir insoluble (Rothenberg 1998, p. 13).

Certains spécialistes ont prudemment proposé des musées d'ordinateurs, où seraient préservées les machines permettant de lire les anciens supports et les logiciels associés (Swade 1998). Toutefois, J. Rothenberg y voit plusieurs problèmes (Rothenberg 1998, p. 18; 1999, pp. 12-13). D'une part, les composants électroniques ont une durée de vie limitée en raison de processus physico-chimiques et leur utilisation accélère ce processus. D'autre part, les fichiers ont toutes les chances d'avoir été migrés sur un autre support non compatible avec l'ordinateur original : il faudrait donc construire de nouvelles interfaces pour permettre la lecture des supports. Dans l'ensemble, les coûts semblent prohibitifs et seuls quelques sites dans le monde pourraient être établis, limitant d'autant l'accès aux documents.

4.3.2 ADN

L'ADN pose la problématique dans un cadre différent. Comme nous l'avons indiqué dans la section 2.2, l'ADN, en tant que base de la vie actuelle sur Terre, est un support qui ne risque pas de devenir obsolète, éliminant ainsi les craintes de J. Mallinson (1986). Cela ne signifie pas pour autant qu'aucune mesure ne doit être prise pour assurer sa préservation à long terme. Il ne s'agit pas ici d'aborder les mécanismes pour la préservation physique de l'ADN (à ce sujet, voir Doricchi et al. 2022, tab. 1), mais bien sa préservation logique en tant que support d'information. Deux axes peuvent s'entrevoir : la transmission de la nature même de l'ADN comme vecteur d'information et la transmission des moyens de décoder l'information.

4.3.2.1 Préservation de la nature de l'ADN comme vecteur d'information

Ce premier axe est relativement peu étudié. On peut relever l'article de Smith et al. (2003, p. 1127), pour qui le codec utilisé pour encoder l'information doit créer une suite de bases qui rendront immédiatement perceptible la nature synthétique des brins d'ADN, révélant ainsi qu'ils ont été créés à dessein. Pour Goldman et al. (2013, p. 77), la structure même des brins d'ADN, de longueur uniforme et sans homopolymères, rendra clair que de tels brins ne sont pas organiques. Toutefois, ce n'est sans doute qu'une question de temps avant que ne se développent de nouvelles techniques de synthèse de l'ADN qui permettront de surmonter ces limitations. Quand chaque fichier, peu importe sa taille, pourra être encodé sur un seul brin, comment s'assurer qu'il soit correctement interprété comme un vecteur d'information numérique et non pas génétique, transmettant, par exemple, les gènes d'une espèce éteinte ? Il sera donc sans doute nécessaire d'imaginer des mécanismes permettant de rendre

perceptible cette nature synthétique lorsque les avancées technologiques suffisantes auront été faites.

La préservation de la nature de l'ADN comme vecteur d'information est en réalité l'étape logique suivant la préservation de la nature de l'archive que nous avons évoquée dans la partie précédente : dans l'éventualité de la perte totale de la documentation associée à l'archive, une conception délibérément artificielle permettrait de faire allusion au caractère informationnel, ou en tous cas non génétique, de l'ADN. À cet effet, l'usage d'origami ADN pourrait offrir des pistes de réflexion intéressantes (Zhan et al. 2023), bien qu'à notre connaissance l'idée n'a jamais été avancée dans la littérature.

4.3.2.2 Préservation des moyens de décoder l'information

Le second axe a naturellement fait l'objet de plus de recherches. Dès le début des années 2000, plusieurs chercheurs s'intéressent à cette problématique en considérant l'ADN comme vecteur d'information textuelle. Ainsi, Cox (2001) a comparé le déchiffrement d'un message encodé sur ADN à celui des langues anciennes comme les hiéroglyphes, soulignant qu'un codec trop complexe limitera son interprétabilité future. Il pointe que le choix de la langue du message encodé n'aurait pas spécialement d'importance, car tout langage dispose d'une structure qui deviendrait visible une fois encodée dans l'ADN ; il encourage toutefois la présence de multiples langages, qui faciliterait le travail de déchiffrement. Enfin, il suggère qu'un langage inconnu pourrait être déchiffré s'il était placé dans un contexte où un sens pourrait être assigné aux mots, que ce soit par les mathématiques, par la logique ou par des images, contexte qui serait préférablement transmis par une clé de déchiffrement qui accompagnerait le message. La même année, Bancroft et al. (2001) posent comme principe que la récupération des données stockées sur ADN « should ideally require minimal prior knowledge beyond a familiarity with molecular biological technique ». Une table de concordance permet de connaître les différents codons, tandis que l'ordre des oligonucléotides est spécifié par un brin d'ADN spécial concaténant toutes les amorces des brins, séparées par une courte chaîne de nucléotides faisant office de séparation. Deux ans plus tard, Smith et al. (2003, p. 1127) soulignent que les brins d'ADN devraient être accompagnés d'un cadre de lecture.

Toutefois, les avancées de Church et al. (2012) et de Goldman et al. (2013) créent un nouveau paradigme et ce sont désormais des données binaires qui sont encodées dans l'ADN, ce qui fait craindre à A. O'Driscoll et R. D. Sleator (2013, p. 125) des difficultés supplémentaires pour la création de « pierres de Rosette moléculaires » capables de garantir le déchiffrement des messages sur des millénaires. En réalité, la translittération des bits à l'ADN n'est guère différente de celle du texte à l'ADN, bien qu'elle soit en général plus complexe pour tenir compte des contraintes biologiques propres à l'ADN, tels que les homopolymères, la proportion de G et de C ou la répétition de motifs. R. R. Garafutdinov et al. (2022) proposent une revue des différents codecs qui ont été proposés dans la littérature, mais ne commentent pas sur leur préservation à long terme.

Les développements du stockage ADN depuis les années 2010 rendent surtout clair que le codec n'est pas l'unique élément à conserver pour permettre la récupération des données. D. Andriamahady, dans son étude sur les moyens pour rendre l'archivage sur ADN compatible avec l'OAIS, liste ainsi les différents documents à conserver vis-à-vis du support (Andriamahady 2021, pp. 47-48) :

- Dictionnaire bit vers ADN
- Données aléatoires pour le masque
- Paramètres des algorithmes Reed-Solomon entrelacés (2 jeux de paramètres)
- Longueur des données pour chaque brin
- Longueur de l'amorce
- Longueur de l'espace pour stocker les identificateurs

Dans l'industrie, deux standards émis par la Storage Networking Industry Association (SNIA) au nom de la DNA Data Storage Alliance proposent d'implémenter une approche par secteur, permettant de graduellement accéder à l'information (SNIA 2023a; 2023b). Le secteur zéro définit l'identité du producteur de l'archive (appelé « vendeur ») et le codec permettant de lire les brins d'ADN du secteur un. Celui-ci contient quant à lui des métadonnées nécessaires à la lecture du contenu de l'archive. Ces deux secteurs ont des présupposés communs :

1. Chaque brin d'ADN contient 150 bases ;
2. Les brins du secteur ne sont composés que des bases naturelles (ACGT) ;
3. Les amorces des secteurs zéro et un sont universellement uniques et ne sont pas utilisées ailleurs dans l'archive ;
4. Sur les 150 bases, 20 sont utilisées de chaque côté du brin pour des amorces génériques et 20 autres sont dédiées aux amorces spécifiques du secteur concerné. Les 70 bases restantes constituent le contenu du secteur, suivant une structure définie dans le standard.

Le secteur zéro présuppose également que :

1. Le secteur zéro tient dans un seul brin ;
2. La table de correspondance pour le vendeur et le codec est documentée et accessible ;
3. Le secteur zéro n'est pas obligatoire pour accéder à l'archive si le producteur de l'archive et le codec sont connus. Toutefois, dans le cas contraire, le secteur zéro permet de les connaître. Le standard recommande l'inclusion du secteur zéro dans tous les cas.

Le secteur un présuppose également que :

1. Le secteur un tient dans plusieurs brins ;
2. Il suit le schéma et la structure définie dans le standard.
3. Son contenu suit la syntaxe du JSON, en excluant les valeurs NULL et est compacté.

Le secteur un définit les métadonnées selon trois niveaux : requis, optionnel et à la discrétion du vendeur. Si l'on s'intéresse aux métadonnées requises, celles-ci sont : l'identifiant de l'archive, les détails du séquençage, les détails du codec et l'identifiant de son vendeur. En outre, les détails du séquençage doivent contenir les informations suivantes : la taille minimale, médiane et maximale des oligonucléotides, le nombre minimal et maximal de réplication d'un oligonucléotide, le nombre d'oligonucléotides, la présence d'une amorce générique et, le cas échéant, l'amorce générique de tête et de queue, la présence de bases naturelles et des renseignements sur la couverture de ces métadonnées. Les détails du codec doivent inclure

quant à eux l'identifiant du codec, son URI et ses paramètres, tandis que son vendeur, son nom et sa version ne sont qu'optionnels.

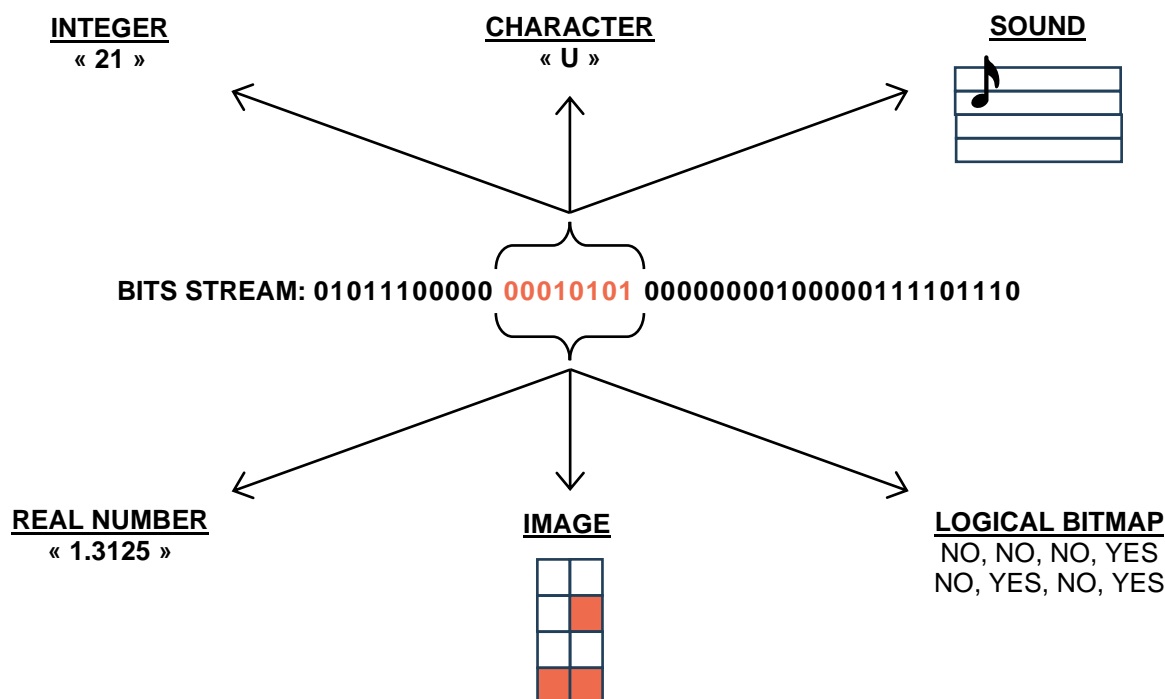
Le standard définissant le secteur un apporte également d'importantes précisions sur le fait que le secteur un doit être lisible depuis l'archive, mais peut également l'être hors de l'archive. Conséquemment, le secteur un peut également être accessible par d'autres moyens que le séquençage, tels que des moyens optiques comme les QR codes, mais aussi d'autres technologies comme les *near-field communication*.

4.4 Préservation des données numériques

4.4.1 Introduction

Par essence, les données numériques ne sont qu'une suite d'uns et de zéros, des bits. Pour accéder à l'information, il est donc nécessaire de correctement interpréter cette suite. Or, comme le montre la Figure 5 ci-après, une suite donnée peut être interprétée d'une multitude de manières. En outre, les bits sont organisés en unités, les bytes, dont la longueur doit également être connue pour permettre la bonne interprétation du document. Enfin, l'information contenue dans les bytes peut également être une référence nécessitant un logiciel particulier pour être correctement interprétée.

Figure 5. Interprétations possibles d'une suite de 8 bits



(adapté de Rothenberg 1995, p. 47)

Dès les années 1990, plusieurs spécialistes voient dans cette réalité un problème plus important que la préservation du support de stockage, mais divergent sur les solutions : un rapport de la *Task Force on Digital Archiving* capitalise sur plusieurs décennies de pratique et prescrit la migration des documents d'un format à l'autre, tandis que J. Rothenberg popularise l'idée d'émuler le logiciel et le matériel permettant d'accéder aux données dans leur suite de bits originelles (Waters, Garrett 1996; Rothenberg 1995; réflexion poursuivie et étendue dans les versions de 1998; et de 1999).

Pour V. Joguin, la migration et l'évaluation sont des approches technologiques plutôt que de véritables stratégies de préservation à long terme, qui, selon lui, devraient être conçues en fonction de la fréquence d'accès aux documents. Il en identifie deux (Joguin 2019, pp. 354-355). La première, qu'il appelle « préservation numérique active », est adaptée à des documents fréquemment lus ; elle fonctionne principalement en maintenant constamment les documents dans un état immédiatement lisible au sein d'un environnement numérique contemporain, principalement en migrant les formats de fichiers. La seconde, que V. Joguin appelle par opposition « préservation numérique passive », concerne des documents dont l'accès est au mieux rare, au pire inexistant, mais dont la conservation est tout de même nécessaire. Cette stratégie se base principalement sur la création d'un environnement numérique immuable, allant d'une certaine manière plus loin que l'émulation. Ces réflexions font écho aux « modes de transmission » évoqués par un rapport de la *Nuclear Energy Agency* (NEA) sur la gestion à long terme des déchets radioactifs, qui fait la distinction entre une transmission de la mémoire par « lien intermédiaire » et par « lien direct » (ou « non-intermédiaire »). Un lien intermédiaire relie le présent et le futur lointain par une chaîne de transferts successifs de documents archivés et d'autres formes de connaissances dans la société, tandis qu'un lien direct s'appuie sur un élément intangible pour relier le présent et le futur lointain (Nuclear Energy Agency 2019a, pp. 54-56).

De manière générale, la nécessité que les migrations soient conduites à intervalles plus ou moins réguliers nous semble en contradiction avec l'objectif de supports de stockage à long terme tels que l'ADN. Naturellement, l'implémentation de standards peut prolonger l'écart entre ces intervalles : toutefois, un rapide survol des recommandations en termes de formats de plusieurs bibliothèques et archives nous apprend que, malgré une tendance de fond certaine pour des formats libres et ouverts, il n'y a pas réellement de consensus sur les formats préférés ou acceptés par type de document (UK Data Service 2022; ETH Zurich 2024; Library of Congress 2024; Bibliothèque nationale de France 2024). Il n'existe par ailleurs aucune garantie que le standard choisi au moment de l'archivage survive au support et, dans le cas de durée de vie de plusieurs centaines d'années, cela est même assez peu probable. J. Rothenberg n'écarterait toutefois pas d'utiliser les standards pour transmettre des métadonnées (Rothenberg 1999, p. 12).

Nous soulignons que nous n'avons pas ici l'ambition de traiter l'ensemble des solutions envisagées pour la préservation à long terme des données numériques, car cela nous emmènerait trop loin par rapport au sujet de ce mémoire. Néanmoins, le développement de l'ADN comme support de stockage à long terme crédible invite à quelques réflexions sur ces aspects. Dans la présente partie, nous nous concentrerons sur l'approche de la préservation numérique à travers et au-delà de l'émulation, car elle nous paraît la plus adaptée au paradigme du stockage sur ADN.

4.4.2 Émulation

J. Rothenberg est un fervent défenseur de l'émulation. Selon lui, les migrations sont laborieuses, coûteuses en temps et en ressources, sujettes aux erreurs et comportent de nombreux risques de perte d'information, ce d'autant plus que l'original n'étant pas conservé il n'est pas possible de savoir ce qui a été perdu au cours de la migration (Rothenberg 1999, pp. 13-16). Il lui apparaît donc nécessaire de s'efforcer de préserver la suite de bits originelle et les moyens de l'interpréter pour éviter toute perte dommageable. À cet effet, il note que :

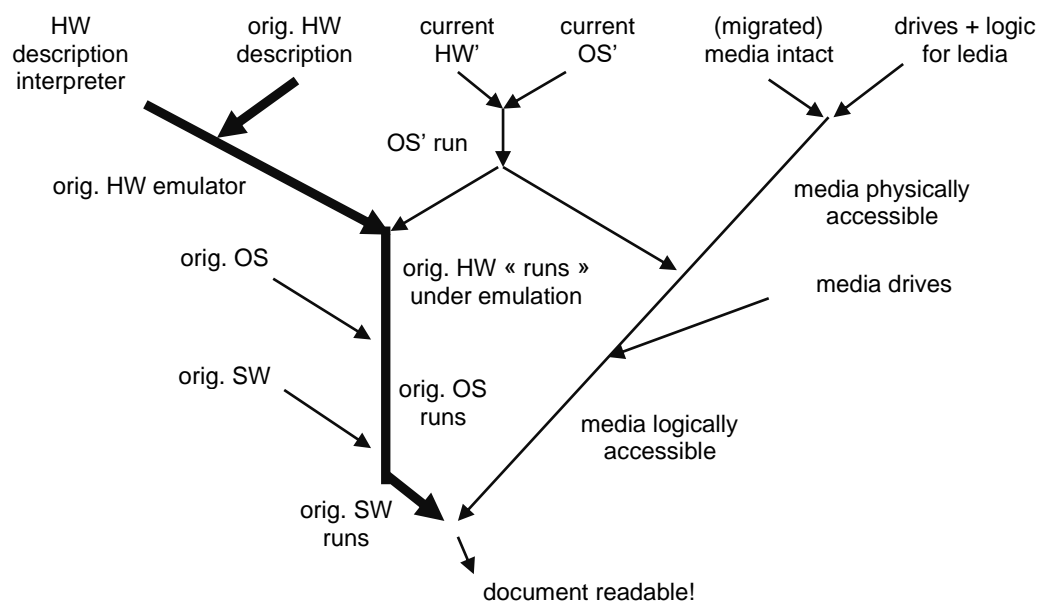
« [I]nterpreting a bit stream depends on knowing how it has been encoded, and a bit stream cannot be fully self-describing, since any description that we encode in the bit stream must itself be interpreted. The only way to bootstrap this process is to include easily-readable annotation with every digital document, explaining how to interpret its bits. » (Rothenberg 1998, p. 19)

Toutefois, pour permettre à ces annotations de survivre aux copies des suites de bits sur des supports différents, J. Rothenberg poursuit :

« [A]ny annotation must itself be stored digitally, along with its associated bit stream; but it must be encoded in a digital form that is more readable than the bit stream itself, in order to serve as a bootstrap. » (Rothenberg 1998, p. 19)

Cette approche rend nécessaire l'encodage des annotations sur un standard contemporain, malgré les défauts que J. Rothenberg leur reconnaît ; mais leur migration est moins problématique, car seule la transmission de l'information sémantique est pertinente. Il faut toutefois que ces annotations puissent être lisibles sans logiciel particulier. La transmission des suites de bits originelles doit concerner non seulement les données, mais également les logiciels permettant de les lire, de même que les informations relatives aux systèmes et aux matériels nécessaire à ces logiciels, afin de pouvoir les émuler dans n'importe quel environnement.

Figure 6. Schéma des prérequis nécessaires à l'émulation



(adapté de Rothenberg 1998, fig. 9)

Se basant sur les travaux de J. Rothenberg, A. R. Heminger et S. Robertson ont proposé au début des 2000 le modèle Digital Rosetta Stone pour la conservation à long terme des documents numériques, dans une perspective inspirée par l'émulation. Ce modèle se base sur trois processus : la préservation des connaissances, la récupération des données et la reconstruction du document (Heminger, Robertson 2000, p. 10). Sur la préservation des connaissances en particulier, la sauvegarde des formats et des logiciels nécessaires pour interpréter les documents, est permise par une « archive de métaconnaissance » (*metaknowledge archive (MKA)*) (Heminger, Robertson 2000, pp. 10-17). Dans leur vision, ce modèle est implémenté non par tout producteur d'archives mais par un organisme dédié qui aurait la responsabilité d'acquérir et de maintenir ces connaissances (Heminger, Robertson

2000, pp. 4-5). Toutefois, une évaluation par méthode Delphi a montré les limites de ce modèle (Heminger, Kelley 2005).

L'approche de J. Rothenberg a suscité des critiques, qui ont en particulier noté le caractère exagéré du modèle qu'il propose (Bearman 1999; Lorie 2001). D. Bearman et R. A. Lorie relèvent en particulier qu'il n'est pas nécessaire de conserver l'environnement de création du document pour en interpréter correctement le contenu. Plus grave, et comme déjà relevé par J. Rothenberg lui-même (Rothenberg 1999, pp. 21-22), préserver l'environnement matériel et logiciel se heurte très probablement à des questions insurmontables de propriété intellectuelle. Enfin, il est avancé que reconstruire un environnement sur la seule base de sa description est loin d'être évident et qu'il sera impossible de savoir s'il fonctionne correctement en l'absence de comparaison avec un système existant. Toutefois, là où D. Bearman défend la migration des documents comme une approche valable, R. A. Lorie propose une nouvelle solution basée sur un ordinateur virtuel universel (*Universal Virtual Computer (UVC)*).

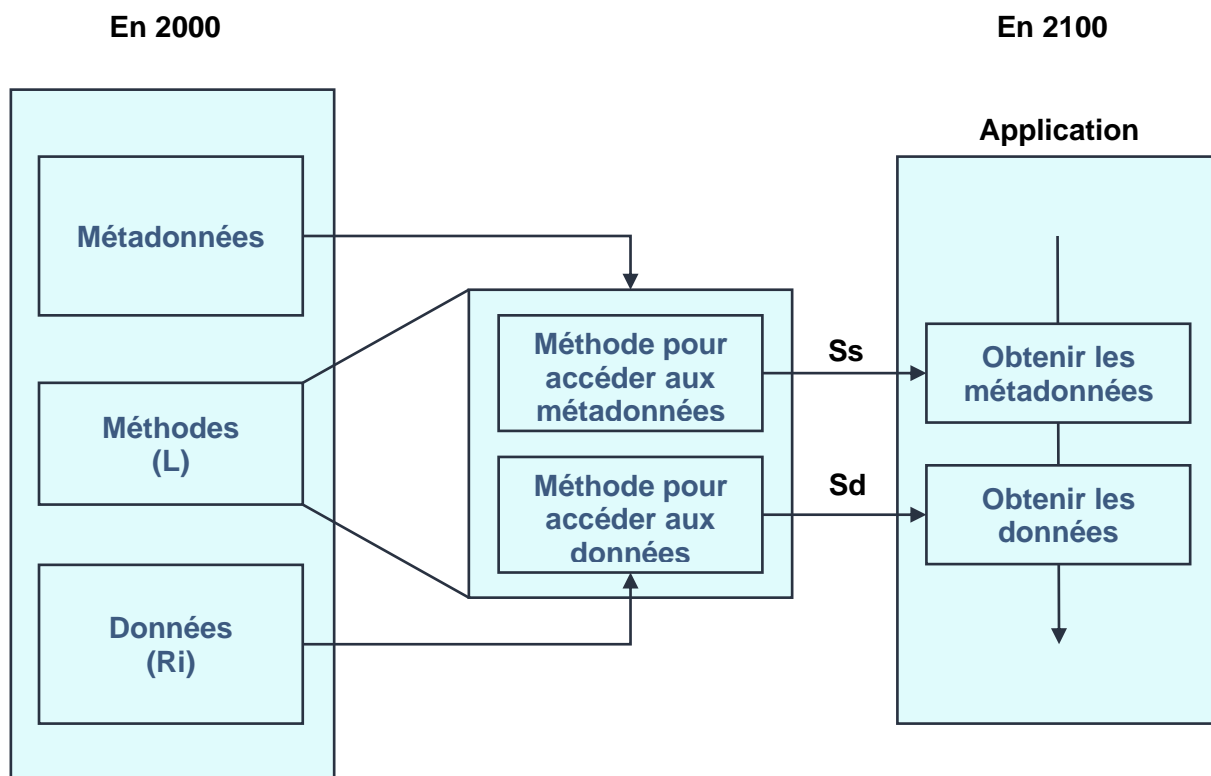
4.4.3 Machine virtuelle

L'approche de R. A. Lorie est guidée par la différenciation entre l'archivage des programmes, qui nécessite une émulation complète, et l'archivage des données qui permet de se contenter d'une émulation partielle (Lorie 2001, p. 347). R. A. Lorie spécifie sa méthode ainsi :

« For data archival, we propose to save a program P that can extract the data from the bit stream and return it to the caller in an understandable way, so that it may be transferred to a new system. The proposal includes a way to specify such a program, based on a Universal Virtual Computer (UVC). To be understandable, the data is returned with additional information, according to the metadata (which is also archived with the data). » (Lorie 2001, pp. 347-348)

Dans le modèle de R. A. Lorie, les données à préserver, organisées selon une représentation donnée (R_i), sont archivées avec leurs métadonnées ainsi qu'avec deux « méthodes » (*methods*), permettant l'une d'extraire les métadonnées (S_s), comprises comme étant la structure des données, et l'autre les données (S_d). L'une et l'autre méthodes doivent être simples, pour minimiser leur taille tout comme leur complexité. Ces deux méthodes sont écrites dans un langage L : pour le choix de L , R. A. Lorie écarte successivement l'usage du langage naturel, d'un langage de haut niveau ainsi que d'un langage informatique particulier pour finalement proposer l'utilisation d'un UVC, supposément indépendant de toute architecture et réalisable dans n'importe quel langage informatique. Finalement, ces deux méthodes sont décrites selon une approche basée sur les schémas DTD utilisés en XML.

Figure 7. Mécanisme général de l'archivage des données selon R. A. Lorie



(adapté de Lorie 2001, fig. 1)

L'idée générale de R. A. Lorie est de remplacer tous les standards de description des données par un unique standard, celui de l'*Universal Virtual Computer*. Cependant, cela nécessite bien évidemment une adoption généralisée de l'UVC pour fonctionner ; à notre connaissance et malgré une implémentation à la Bibliothèque Nationale des Pays-Bas (Lorie, van Diessen 2005; van der Hoeven, van Diessen, van der Meer 2005), cela n'a pas eu lieu. Toutefois, l'idée de R. A. Lorie a fait des émules et dans les années suivantes plusieurs projets ont opté pour la machine virtuelle comme méthode de préservation des documents numériques.

Parmi ceux-ci, le modèle Cuneiform nous semble particulièrement intéressant entre autres pour son ambition d'être implémentable lors d'un « fun afternoon hack » (Nguyen, Kay 2015). Ses auteurs reprochent en effet à l'UVC précisément sa vocation universelle, qui selon eux le rendra inévitablement compliqué s'il doit être compatible avec toutes les architectures possibles : il en résulte une documentation de plusieurs dizaines de page et un temps d'implémentation estimé à plusieurs semaines. En comparaison, le modèle Cuneiform se veut beaucoup plus simple. Il se compose de trois parties principales : l'étiquette, l'en-tête et le programme lui-même. L'étiquette, conçue pour être lisible à l'œil nu, contient une brève description du programme conservé, des instructions sur la lecture des données depuis le support de stockage et des instructions sur l'interprétation de la suite de bits. L'en-tête, qui précède le programme, est un bitmap en noir et blanc contenant un document qui explique comment interpréter le reste du flux de bits. Le programme est le logiciel à préserver, encapsulé dans cette suite de bits structurée. Pour faciliter la préservation et l'émulation des programmes, le modèle Cuneiform intègre une machine virtuelle minimaliste nommée Chifir, suffisamment simple pour être décrite en une seule page et implémentée en une seule après-midi. Elle utilise un jeu d'instructions basique et une architecture de mémoire simple pour garantir qu'elle puisse être reproduite avec des connaissances et des technologies futures.

La méthode fonctionne comme suit : dans le présent, un programme est écrit pour une plateforme informatique spécifique. Lors de la préservation, un émulateur pour cette plateforme originale est créé pour fonctionner sur Chifir. Cet émulateur est ensuite encapsulé avec le programme et stocké. Dans le futur, un archéologue peut lire la spécification de Chifir contenue dans l'en-tête et implémenter un émulateur pour cette machine virtuelle simple. Ensuite, il peut utiliser cet émulateur pour exécuter l'émulateur de la plateforme originale et enfin le programme conservé. Malgré un certain nombre de critiques qu'il lui adresse, K. J. Sitaker décrit Chifir comme « the first archival virtual machine good enough to criticize » (Sitaker 2018).

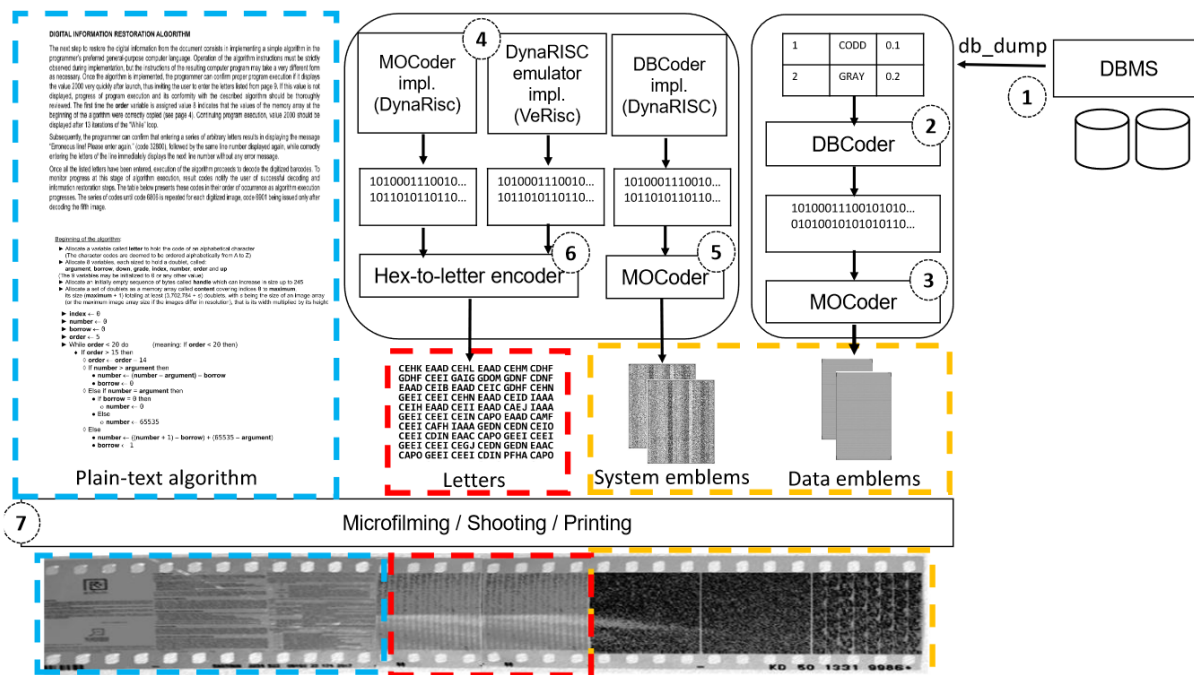
En France, la société Eupalia a développé la solution Micr'Olonys, une stratégie de « préservation numérique passive » (Joguin 2019; Appuswamy, Joguin 2021). Cette solution repose sur trois éléments : DBCoder, qui permet d'encoder et de décoder les données binaires avec une forte compression ; MOCoder, qui permet d'encoder et de décoder les données binaires sur un format analogique ; et enfin Olonys, un émulateur universel. Les modules d'encodage des éléments DBCoder et MOCoder sont écrits en C# et sont conçus pour être exécutés dans le présent. À l'inverse, les modules de décodages sont prévus pour être exécutés dans un temps indéterminé et pour cette raison doivent être écrits de manière à être implémentés indépendamment des moyens technologiques futurs. Eupalia a donc pris le parti de créer Olonys, un émulateur simulant un processeur RISC-V (*Reduced Instruction Set Computer Five*) simplifié, appelé DynaRisc, permettant d'exécuter 23 opérations. De plus, Olonys simule également un second processeur RISC de seulement 4 opérations, appelé VeRisc, qui sera chargé d'émuler DynaRisc. Ainsi, un utilisateur peut se contenter d'écrire un émulateur pour VeRisc pour charger l'émulateur DynaRisc qui, à son tour, exécute les instructions des décodeurs.

Concrètement, la solution Micr'Olonys fonctionne en sept étapes réparties dans trois ensembles. Dans le premier ensemble, Micr'Olonys procède à l'archivage des données en les extrayant tout d'abord du système original. Ensuite, les données sont compressées par DBCoder. Enfin, elles sont transformées en code-barres 2D, appelés « emblèmes », pour former des « emblèmes-données ».

Dans le deuxième ensemble, Micr'Olonys procède à l'archivage des décodeurs, écrits avec DynaRisc, ainsi qu'à l'émulateur VeRisc. Tout d'abord, la quatrième étape consiste à écrire les instructions des décodeurs de DBCoder et de MOCoder en langage DynaRisc. Ensuite, Micr'Olonys convertit les instructions de DBCoder en « emblèmes-système ». Les instructions de MOCoder et de l'émulateur DynaRisc, qui sont nécessaires afin de pouvoir lire les emblèmes, ne peuvent quant à elles pas être converties en emblèmes. Micr'Olonys procède donc à la conversion de leurs valeurs binaires en valeurs hexadécimales, qui sont encodées avec les lettres A à P pour représenter les valeurs 0xF à 0x0. Ces valeurs sont accompagnées d'un pseudo-code et d'instructions pour permettre d'implémenter VeRisc, le tout formant le bootstrap initial pour le lancement du décodage.

Enfin, la septième étape consiste à écrire le bootstrap, les « emblèmes-système » et les « emblèmes-données » soit sur papier permanent (ISO 1994) soit sur microfilm, deux supports analogiques dont la durée de vie est d'au moins 500 ans.

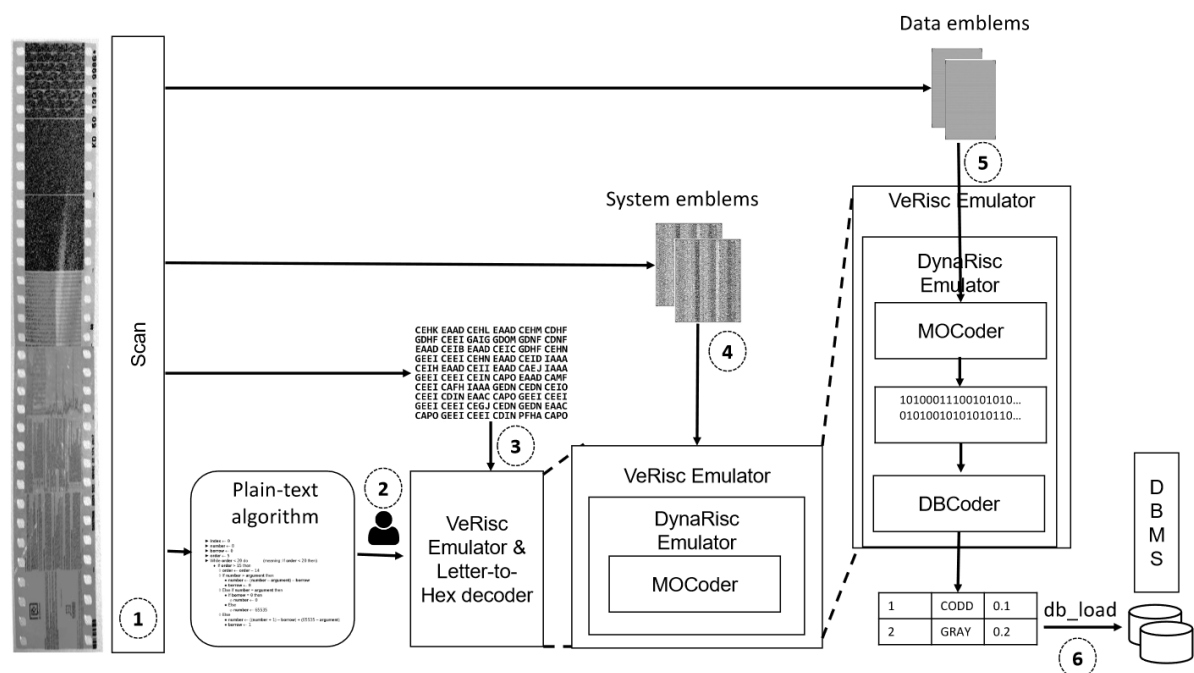
Figure 8. Étapes d'encodage de la solution Micr'Olonys



(Appuswamy, Joquin 2021, fig. 2)

Pour pouvoir décoder, un utilisateur doit commencer par scanner l'entièreté du document et prétraiter les images selon certaines indications du bootstrap. Il implémente ensuite l'émulateur VeRisc sur la base des instructions du bootstrap et l'exécute. Le reste est ensuite géré par Micr'Olonys.

Figure 9. Étapes de décodage de la solution Micr'Olonys



(Appuswamy, Joguin 2021, fig. 3)

L'ANDRA a procédé à un test pour vérifier l'utilisabilité de la solution (Joguin, Dumont 2022). Deux informaticiens, dont un étudiant en formation, ont réussi à faire sens du document et à

restaurer la base de données sans connaissances préalables de la solution Micr'Olonys. Ce test ayant été concluant, l'ANDRA a procédé à la transcription d'une base de données contenant 1,5 million de lignes en un fichier de 464 pages, y compris les instructions, alors qu'il aurait fallu plus d'un million de pages pour retranscrire la base de données en plein texte. Certaines spécificités doivent être relevées : premièrement, les instructions ont été écrites en deux langues, en français et en anglais, afin de maximiser les chances de leur compréhension dans le futur, et ont utilisé un langage simplifié. Deuxièmement, la base de données a été extraite en CSV, un format de fichier simple et répandu. Troisièmement, les spécifications du format ont également été incluses, de même qu'une table UTF-8 des caractères présents dans le document. Enfin, il est souligné que les données devraient être placées dans un contexte permettant de décrire ce qu'elles contiennent, comment elles devraient être utilisées, leur relation avec d'autres documents, etc.

En revanche, comme noté par ces auteurs à plusieurs reprises, la densité d'information qu'il est possible de stocker sur papier et sur microfilm est relativement faible par rapports aux supports de stockage numériques. Si la réduction du volume à imprimer telle que démontrée reste impressionnante, la base de données initiale ne mesurait que 626 MB, bien loin des volumes de données actuellement générés. La méthode Micr'Olonys ne saurait donc être applicable au stockage d'un nombre important de documents. Conscients de ce problème, Appuswamy et Joguín (2021, p. 7) annoncent travailler à une conjonction de Micr'Olonys avec un stockage de l'ADN, qui permettrait de stocker les données sur ce support tandis que les décodeurs resteraient stockés sur des supports analogiques. De plus, comme l'a reconnu V. Joguín lui-même lors de sa présentation à l'iPres 2022 à la suite d'une question d'une auditrice, il serait sans doute nécessaire de transmettre également les instructions pour la construction d'un ordinateur et d'un scanner afin de fournir tous les éléments pour accéder aux données.

Toutefois, il n'est pas clair si Micr'Olonys fonctionne avec des formats non textuels. Parallèlement, une autre équipe de chercheurs a également travaillé à l'implémentation d'une solution pour permettre la préservation à long terme des données numériques dans leur format d'origine, en supportant pour l'instant le JPEG, le TIFF et le PDF/A (Rummelhoff et al. 2021). Cette solution prend la forme d'un compilateur C, d'une « machine abstraite » permettant d'implémenter ce compilateur et d'instructions permettant d'implémenter la « machine abstraite ». Ces instructions sont produites en trois versions alternatives avec différentes audiences en tête, dont deux visent un public dans un futur indéterminé. Selon les chercheurs, cette solution présente deux avantages, en permettant de s'affranchir de tout logiciel pour interpréter les données et en évitant d'expliquer les formats. Enfin, mentionnons également la solution éponyme développée par l'entreprise Piql et dont le fonctionnement semble proche de celui de Micr'Olonys (Sabliński, Trujillo 2021).

4.4.4 ADN

Dans le cadre du stockage sur ADN, plusieurs études ont proposé des mécanismes originaux pour préserver les métadonnées. Ainsi, Bancroft et al. (2001), avant la généralisation de l'index placé avec la charge utile, suggère l'utilisation d'une « clé polyamorce » (*polyprimer key* (PPK)) pour stocker la liste des amorces des différents brins d'ADN ; cette PPK est en réalité un unique oligonucléotide constitué des différentes amorces mises bout à bout et séparées par une courte chaîne de nucléotides faisant office de séparation. Dans l'optique de Bancroft et al., la PPK est le premier élément amplifié dans l'archive ADN. De leur côté, M. Ailenberg

et O. D. Rotstein (2009) proposent qu'un index constitué de plasmides puisse contenir des informations générales sur l'archive, dont le titre, la date, les auteurs, la taille du contenu et le nombre de plasmides.

À notre connaissance, une seule étude (Li et al. 2021) s'est intéressée à préserver sur un même support de stockage ADN les logiciels utiles à l'interprétation des données, ce dans une optique d'autonomie et d'auto-description et avec l'intention claire de ne pas introduire de dépendances externes. L'objectif de cette étude est de pouvoir transmettre des logiciels de compression en même temps que les données compressées, afin de permettre une décompression pérenne des données.

Afin d'y parvenir, l'étude présente d'abord quatre méthodes d'autonomie. La première, intitulée « méthode intuitive », consiste à englober dans un seul et même paquet le logiciel et les différents fichiers qui ont besoin de ce logiciel. Comme noté par les chercheurs, cette méthode semble toutefois peu adaptée aux spécificités du stockage sur ADN, puisque tout un paquet devrait être lu et séquencé avant de pouvoir accéder à un unique fichier d'intérêt, ce qui avec les technologies actuelles nécessiterait beaucoup de temps et des coûts importants. Pour répondre à cela, une seconde méthode, intitulée *1-1CS (1-1 Continuous Storage)*, est envisagée, dans laquelle les fichiers sont empaquetés indépendamment les uns des autres, mais chacun avec le logiciel nécessaire à leur lecture. Cette méthode permet d'accéder facilement à n'importe quel fichier, mais au prix d'importantes pertes de densité dans le stockage, puisque les données de chaque outil sont dupliquées pour chaque fichier qui l'utilise (Li et al. 2021, pp. 3, 5).

Pour résoudre ce problème tout en conservant l'accès direct à chaque fichier, les chercheurs ont proposé deux autres méthodes, *M-1CI (Many-to-One Chain Indexing)* et *1-MCI (One-to-Many Chain Indexing)*, aux principes similaires. Ces deux méthodes font usage des amorces ADN pour relier les fichiers et les outils, qui sont tous empaquetés indépendamment. La méthode M-1CI consiste à associer à chaque fichier l'amorce de l'outil ; l'amorce est exprimée directement sur le dernier oligonucléotide du fichier. En tant que telle, l'amorce de l'outil n'est donc accessible qu'après avoir séquencé et recomposé le fichier. Le désavantage d'une telle approche, comme noté par Li et al., est qu'il rend nécessaire d'avoir recours à deux PCRs successives, la première pour le fichier, la seconde pour l'outil associé (Li et al. 2021, pp. 4-5). Pour répondre à cela, l'approche 1-MCI a été développée, où chaque oligonucléotide encodant l'outil contient l'amorce des fichiers qui l'utilisent, permettant à une seule PCR d'amplifier l'outil et le fichier associé. Toutefois, considérant la limite de deux cents bases par brin pour la synthèse, l'approche 1-MCI semble impraticable dès lors qu'une dizaine de fichiers utilisent le même outil, car un même brin ne pourrait contenir toutes leurs amorces (Li et al. 2021, pp. 4-6).

Comme signalé par Li et al. (2021, pp. 6-7), les méthodes d'autonomie présentées ci-dessus nécessitent également des mécanismes d'auto-description. Ceux-ci portent à la fois sur la structure du fichier encodé ainsi que sur la structure du brin. Cette dernière est relativement standard, avec des amorces de chaque côté, un index, la charge utile, le code de correction et l'indication de s'il s'agit d'un fichier ou d'un outil. La structure du fichier est plus intéressante : elle est divisée en plusieurs champs, permettant d'indiquer le type du fichier, son identifiant, sa méthode de stockage, la longueur des données, les données elles-mêmes et le cas échéant l'outil utilisé. La disponibilité de certains champs variant selon le contenu d'autres, la manière de lire les différents champs peut facilement être naviguée à l'aide d'un algorithme simple, transmis par pseudo-code (Li et al. 2021, p. 7).

5. Élaboration des spécificités d'un connecteur ADN

5.1 Introduction

Les chapitres précédents ont introduit les concepts du stockage sur ADN et du modèle OAIS ainsi qu'une revue de la littérature sur les mécanismes de préservation à long terme autour de trois axes : la préservation de la partie immédiatement perceptible de l'archive, la préservation logique du support et la préservation des données. Il s'agit maintenant d'en faire la synthèse. L'objectif de ce mémoire étant de définir les spécificités d'un connecteur pour archiver des données numériques sur ADN, en prenant en compte les contraintes propres à ce support de stockage et en appliquant les principes du modèle OAIS, nous nous intéresserons plus particulièrement aux possibilités d'auto-description et d'autonomie pour permettre la pérennité d'un Plan de reprise d'activité sur le long terme. Pour le dire différemment, nous proposerons des recommandations pour s'assurer que le contenu des archives stockées sur ADN reste accessible alors que les technologies et la société auront évoluées tout au long de la durée de vie du support de stockage.

Au sein du présent chapitre, nous procéderons de la manière suivante. Nous reviendrons tout d'abord sur les concepts de Plan de reprise d'activité et d'auto-description, afin de clarifier l'usage que nous en faisons. Nous poursuivrons avec les particularités du projet DNAMIC et ses liens avec la technologie DLCM ; nous présenterons également les modalités des deux scénarios envisagés par le projet pour le connecteur ADN. Puis, nous aborderons en premier lieu les spécificités qui nous semble communes aux deux scénarios. Ensuite, nous nous concentrerons sur les spécificités propres à chacun des scénarios.

5.2 Définition des concepts

Nous avons vu dans la section 3.3 que, dans le cadre du modèle OAIS, le Plan de reprise d'activité intervient au sein de l'entité fonctionnelle Stockage, où il y est essentiellement décrit comme une fonction offrant un mécanisme de duplication des collections archivées. Dans cette optique, il répond donc essentiellement à des échecs du support de stockage courant dont, dans l'hypothèse d'un sinistre, le contenu serait tout ou en partie perdu. Cela est illustré par le schéma (Figure 4), où le rôle du Plan de reprise d'activité est essentiellement celui d'un pont entre le ou les supports de stockage courants de l'OAIS et un ou des supports de stockage de sauvegarde.

Cependant, la réplication du support physique de stockage ne peut que répondre à des besoins de préservation à relativement court terme, de l'ordre de quelques années. Un Plan de reprise d'activité de ce type présuppose une continuité de l'organisation responsable de l'archivage ainsi qu'une maîtrise du matériel, des logiciels et des processus employés pour le stockage des données. Le cadre de l'archivage sur ADN, et plus généralement le recours à des supports de stockage à long terme sur plusieurs décennies ou siècles, pose la question différemment, puisque sur ce laps de temps il ne peut exister aucune garantie que l'environnement organisationnel et surtout technique de l'archive reste le même. Dans l'approche qui fait l'objet du présent mémoire, la définition du Plan de reprise d'activité doit donc être adaptée.

À cet effet et dans la suite de ce mémoire, nous considérerons qu'un Plan de reprise d'activité conforme doit non seulement offrir la duplication des collections archivées et ce sur un support stable et pérenne, mais doit également mettre en œuvre les mécanismes nécessaires pour

permettre l'accès à ces collections dans le cas d'une interruption de l'activité sur une durée indéterminée, qui serait suffisamment longue ou grave pour avoir entraîné une perte des connaissances nécessaires à l'accès aux données.

Un tel Plan de reprise d'activité repose nécessairement sur des stratégies d'auto-description (*self-description*), c'est-à-dire la propriété d'un objet à avoir sa description incluse avec lui de telle sorte à ce qu'elle soit toujours accessible. L'expression apparaît à quelques endroits dans le modèle OAIS mais n'est jamais réellement définie et l'auto-description n'est en tout cas pas explicitement requise dans l'application de l'un ou l'autre concept. Elle est pourtant au cœur du modèle OAIS : la structure d'un AIP (Figure 3) montre que tous éléments permettant de l'interpréter, en particulier l'Information de représentation nécessaire à la bonne interprétation des données qu'il contient, sont inclus avec lui. La récursivité de l'Information de représentation jusqu'à un support physique assure également qu'elle puisse être accessible à un être humain sous une forme aussi peu technologiquement dépendante que possible. Dans le cas d'un Plan de reprise d'activité tel que nous l'avons défini, l'auto-description doit couvrir tous les aspects de l'OAIS, des niveaux les plus abstraits aux plus concrets. Dans cette approche, l'auto-description recoupe également l'autonomie (*self-containment*), c'est-à-dire l'absence de dépendance externe au bon fonctionnement d'un objet. Une OAIS auto-décrite et autonome se suffit donc à elle-même pour être compréhensible et interprétable.

5.3 Technologies et scénarios envisagés par DNAMIC

Comme présenté en introduction de ce mémoire, le projet DNAMIC s'intègre avec la technologie DLCM, un système de préservation à long terme pour la sauvegarde des données de la recherche et des données patrimoniales. La technologie DLCM se veut conforme au modèle OAIS. Au sein de la technologie DLCM, un AIP est composé d'un fichier ZIP comprenant les documents (Information de contenu) ainsi que des métadonnées en XML, comprenant les métadonnées descriptives (Information de pérennisation) et les métadonnées administratives (Information de représentation) (Burgi 2024, p. 11).

Dans le cadre de l'intégration du stockage ADN pour la technologie DLCM, plusieurs choix de conception ont déjà été effectués. Les brins d'ADN mesurent deux cents nucléotides, dont vingt à chaque extrémité sont des amorces (Burgi et al. 2022, p. 3). Ces amorces sont générées par un algorithme sur la base de l'identifiant de l'AIP, lui-même basé sur son empreinte numérique (Burgi 2024, p. 11). Outre la charge utile, les brins d'ADN contiennent également un index et des codes de corrections d'erreur Reed-Solomon (Burgi et al. 2022, p. 3).

Le projet DNAMIC envisage deux scénarios pour l'intégration avec la technologie DLCM. Dans le premier scénario, qui représente une version simplifiée, toute la stratégie de stockage est implémentée par des séquences de nucléotides, sans recours aux nanostructures. De plus, les AIPs restent complets et ne sont pas divisés selon leurs différents éléments ; ils sont empaquetés dans un fichier ZIP. Optionnellement, certaines métadonnées peuvent être encodées et stockées physiquement à part (Burgi 2024, pp. 14-15).

Inversement, le second scénario propose une version plus avancée. La stratégie de stockage fait appel tant aux séquences de nucléotides qu'aux nanostructures. Les AIPs peuvent être transformés, notamment pour permettre un encodage fichier par fichier ou pour distinguer les métadonnées des données ; il n'y a pas d'obligation de conserver le tout dans un fichier ZIP. Les métadonnées doivent être encodées et stockées avec les données.

Il est nécessaire de souligner que dans le cadre du projet DNAMIC, il est déjà envisagé que deux codecs distincts puissent être utilisés au sein de la même archive. À notre connaissance, une telle possibilité est sans précédent dans la littérature sur le stockage ADN ; en particulier, les deux standards de la DNA Data Storage Alliance n'en font pas mention. Dans cette situation, il n'est plus suffisant de transmettre la documentation liée aux codecs : il faut également que chaque brin d'ADN puisse être rattaché à un codec afin de savoir quels paramètres et quels algorithmes vont s'appliquer. Cette étape supplémentaire, mais cruciale dans une optique d'auto-description, va nécessiter des approches différentes selon les scénarios.

5.4 Spécificités communes aux deux scénarios

Comme annoncé dans l'introduction de ce chapitre, nous commençons par envisager les spécificités communes aux deux scénarios envisagés. Ceux-ci se différenciant dès le niveau du support, par l'usage exclusifs de brins d'ADN d'une part et par un mélange de brins d'ADN et de nanostructures d'autre part, cette partie se limite à tous les éléments nécessaires et accessibles avant même d'ouvrir le support et d'en extraire les données. En grande partie, il s'agit donc de détailler la documentation nécessaire. Nous ferons également quelques commentaires sur le choix du codec et sur le format des données.

5.4.1 Généralités

Il paraît opportun de commencer cette section par quelques généralités. Tout d'abord, il nous paraît plus pérenne de faire des choix de conception visant à la simplicité, quitte à y sacrifier l'efficacité. Ainsi, le principe de Bancroft et al. (2001) selon qui la récupération des données stockées sur ADN « should ideally require minimal prior knowledge beyond a familiarity with molecular biological technique », de même que celui de N. T. Nguyen et A. Kay (2015) dont la machine virtuelle est implémentable en une « fun afternoon hack », nous semblent être des approches à privilégier dans la conception d'un connecteur ADN. Le choix conscient de la simplicité est un garant d'une transmission plus réussie, en réduisant les efforts nécessaires et en limitant les problèmes potentiels (Joguin 2019, p. 356; Rummelhoff et al. 2021, p. 154917). Il n'y a selon nous aucun intérêt à développer des processus assurant une efficacité maximale si leur documentation couvre des milliers de pages qui rendront leur recréation *ex nihilo* difficile voire impossible.

En outre, le concept de la redondance dans la transmission de l'information, tel qu'il est avancé par la Nuclear Energy Agency (2019a, pp. 22-23), nous paraît judicieux, à tout le moins parce qu'il part du postulat qu'une partie de la documentation puisse disparaître et que l'accès à l'information doit rester possible malgré cela. Dans le cadre de ces recommandations, nous choisissons d'écarter la redondance par la multiplication de la documentation sur plusieurs lieux de stockage, qui nous semble surtout justifiée par l'objet des archives envisagées par la Nuclear Energy Agency ; en outre, il paraît probable qu'en cas de perte totale de la documentation physiquement proche de l'archive, celle-ci soit également perdue. Nous considérerons en revanche les liens entre les documents et l'usage de modes de communication multiples et différents comme pertinents pour la préservation à long terme d'une archive sur ADN.

5.4.2 Documentation analogique nécessaire

Le besoin d'une documentation découle du concept d'Information de représentation tel qu'il est décrit dans le modèle OAIS et selon lequel toute donnée doit être accompagnée

d'information permettant de l'interpréter et ce de manière récursive jusqu'à un support analogique (CCSDS 2019, sect. 4.2.1.3.2). Deux points méritent d'être signalés : d'une part, il ne s'agit pas ici de conserver l'intégralité de la documentation liée à l'archive, dont la masse la rendrait impraticable (Nuclear Energy Agency 2019a, p. 64), mais de sélectionner les documents nécessaires à l'accès aux données. D'autre part, la documentation ne doit pas seulement être préservée de manière indépendante du support ADN, mais bien indépendamment de toute technologie numérique. En outre, afin d'atteindre une forme d'autonomie pour l'archive, l'inclusion de tout matériel pertinent réduira voire éliminera le besoin de recourir à des dépendances externes pour accéder à l'information ; à ce titre, le recours à un organisme externe comme décrit dans le modèle Digital Rosetta Stone, que nous avons décrit à la section 4.4.2, ne nous semble pas envisageable (Heminger, Robertson 2000, pp. 4-5).

5.4.2.1 Principes généraux

5.4.2.1.1 Choix du support analogique

Dans le cadre de ce mémoire et considérant le volume que pourraient représenter les documents ci-après, nous recommandons plutôt l'usage de papier permanent. Celui-ci est prévu pour durer entre 600 et 1000 ans en conditions normales (Dumont, Charton 2012) et cette durée de vie pourrait être étendue si les documents sont scellés dans l'environnement protégé d'une capsule temporelle. Alternativement, si la documentation à conserver est abondante, nous recommandons le microfilm qui offre une meilleure densité. À ce titre, une conception suggérant l'usage de procédés d'agrandissement optique pour accéder au reste de la documentation, tel que proposé par la fondation Long Now (Kelly 2008), paraît pertinent.

5.4.2.1.2 Langage

En suivant en particulier les recommandations de la Nuclear Energy Agency, nous recommandons que toute documentation soit écrite *a minima* dans la langue locale et dans la *lingua franca* actuelle (Nuclear Energy Agency 2019b, pp. 19-21). Si la place le permet, toute autre langue pertinente, qu'elle soit nationale ou d'intérêt pour les données de l'OAIS, doit être considérée. De plus, comme recommandé par le modèle OAIS, toute langue utilisée devrait également être accompagnée d'un dictionnaire et d'une grammaire (CCSDS 2019, sect. 4.2.1.4.1). Tout élément ayant une base algorithmique devrait quand à lui être écrit en pseudo-code afin de permettre sa réalisation dans n'importe quel langage.

5.4.2.1.3 Technicité

Afin de rester cohérent avec le principe de la simplicité que nous avons évoqué précédemment, nous recommandons de conserver un niveau de langage simple et de ne pas présumer des connaissances des lecteurs futurs ; tous les termes techniques utilisés devraient ainsi être définis (Rummelhoff et al. 2021, p. 154916). Dans l'optique de la constitution d'un SER, composé de documents produits dans le cadre des opérations courantes de l'archive, il sera probablement nécessaire de les contextualiser au préalable, voire de les éditer pour en modifier certains termes. Sur certaines parties particulièrement cruciales, où l'économie de termes techniques pourrait ne pas être possible, dupliquer l'information en proposant deux versions, écrites pour des publics cibles différents, pourrait permettre de s'assurer qu'au moins une puisse être comprise (Rummelhoff et al. 2021, p. 154916). De plus, dans la même optique de simplicité, limiter la masse de la documentation en se tenant aux informations nécessaires permettrait de faciliter leur compréhension.

5.4.2.2 Choix des documents

5.4.2.2.1 KIF et SER

L'inclusion en premier lieu d'un document récapitulatif et général sur l'archive, son contexte de création et son contenu, nous semble obligatoire afin de permettre à toute personne qui rentrerait en contact avec l'archive d'agir en toute connaissance de cause et en particulier de décider si elle souhaite décoder l'archive ou bien continuer à la préserver sans la manipuler, par exemple si ses connaissances techniques ne permettent pas de la décoder.

Si un utilisateur décide d'accéder l'archive, le KIF devrait également pointer vers les différents documents décrits dans les parties suivantes. Nous signalons ces documents comme nécessaires sans présumer qu'ils existent au préalable dans le cadre du fonctionnement courant de l'archive. À ce titre, ils ne composent pas tout à fait un SER tel qu'envisagé par la Nuclear Energy Agency (Nuclear Energy Agency 2019c). Il nous semble que c'est ce genre de document qu'envisage D. Andriamahady lorsqu'elle recommande l'inclusion d'une « documentation générale sur les procédures du projet » (Andriamahady 2021, p. 48). La décision d'inclure ou non tel document dans cet ensemble pourrait être prise sur la base du tableau proposé par la Nuclear Energy Agency (voir Tableau 1).

5.4.2.2.2 Explications sur l'ADN et les techniques de séquençage

Ce document sur les techniques de séquençage doit permettre de suppléer la connaissance minimale requise en biologie moléculaire pour comprendre comment sont stockées les données et comment y accéder, tel que recommandé par Bancroft et al. (2001). À la lecture de ce document, l'utilisateur doit pouvoir se faire une idée de la structure physique de l'ADN. Ce document doit également fournir des indications sur les procédures de séquençage afin de permettre la lecture des suites de nucléotides ; dans cette optique, un tel document est prévu par D. Andriamahady (2021, p. 48).

Il ne serait peut-être pas inopportun d'inclure avec ce document un échantillon d'ADN ne contenant pas d'information, afin de permettre à un potentiel utilisateur de se familiariser avec l'ADN sans compromettre les données.

5.4.2.2.3 Paramètres d'encodage et codec

Ces informations ont déjà été identifiées par le projet DNAMIC comme capitales pour assurer la bonne transmission de l'information (Burgi 2024, p. 13). Ce sont également ces informations que les standards de la DNA Data Storage Alliance (SNIA 2023a; 2023b) prévoient de transmettre. En effet, leur perte compromettrait fortement, voire rendrait impossible, le décodage de l'information.

Toutefois, plusieurs points nous semblent problématiques tant dans l'approche actuellement choisie par le projet DNAMIC que par celle des standards de la DNA Data Storage Alliance. En particulier, la transmission de ces informations uniquement via le même support qu'elles sont censées aider à comprendre nous semble dommageable. La pratique existe pour d'autres supports de stockage, notamment optiques (Solomon et al. 2021; Appuswamy, Joguín 2021; Sabliński, Trujillo 2021), mais la documentation est alors explicitement prévue pour être accessible avec un minimum de technologie et sans cryptage. Il nous semble nécessaire que des informations aussi essentielles pour l'accès aux données soient transmises de manière à être immédiatement perceptibles pour un être humain en vue de réaliser une auto-description ;

en d'autres termes, nous rejetons l'idée que ces informations ne puissent être accessibles que par séquençage, tout en n'excluant pas l'idée qu'une copie puisse exister sous cette forme.

Or, comme nous avons pu le voir, le standard du secteur un de la DNA Data Storage Alliance prévoit que les informations qu'il contient puissent être accessibles par d'autres technologies que le séquençage (SNIA 2023b). On peut s'étonner à juste titre que ces précisions ne sont pas présentes pour le standard du secteur zéro, transmettant le codec, où les enjeux sont d'autant plus importants.

Ceci nous semble même aggravé par le fait que le séquençage du secteur zéro ne transmet pas le codec lui-même, mais livre deux identifiants qu'il faut aller ensuite rechercher dans une table de concordance externe accessible uniquement par Internet (SNIA 2023a). Cette approche ne semble pas très différente de la *metaknowledge archive* proposée par le modèle Digital Rosetta Stone, où les informations sur les différents formats de fichiers sont centralisées au sein d'un unique office : l'évaluation DELPHI de ce modèle révèle que le maintien d'une telle base entraînerait des coûts significatifs, tandis que son échec condamnerait l'ensemble du processus (Heminger, Kelley 2005, pp. 17-18, 28). Bien que le maintien de la table de concordance telle que prévue par le standard de la DNA Data Storage Alliance représente une charge moindre, une telle approche n'offre aucune garantie de pérennité et n'est selon nous pas compatible avec une approche basée sur l'autonomie et l'auto-description. En effet, un tel défaut est déjà identifié par le projet DNAMIC, qui relève que « maintaining such a database over the very long term is illusory » (Burgi 2024, p. 13).

Nous recommandons donc d'inclure toute la documentation relative au codec et à la structure des brins sous forme analogique, comme cela est déjà recommandé par D. Andriamahady (Andriamahady 2021, p. 48).

5.4.3 Documentation analogique supplémentaire

La documentation décrite ci-dessus doit permettre la récupération des données. À ce stade, il serait donc en théorie possible de transmettre toute documentation supplémentaire via son encodage sur ADN. Néanmoins, par soucis de redondance comme prévu par la Nuclear Energy Agency, il peut être souhaitable de disposer également de cette documentation sur support analogique.

5.4.4 Choix du codec

Les deux scénarios envisageant d'utiliser des brins d'ADN pour encoder l'information, la question du choix du codec se pose dans les deux cas. Nous avons déjà pu écrire précédemment que, quel que soit le choix du codec, celui-ci devra être soigneusement documenté pour en permettre le décodage. Dans cette optique, il convient de modérer la complexité du codec tant pour éviter que sa documentation ne prenne trop d'ampleur que pour faciliter son implémentation.

Dans la mesure où le projet DNAMIC envisage déjà que plusieurs codecs puissent coexister dans une même archive, nous proposons que lorsque les métadonnées peuvent être séparées des fichiers ZIP empaquetant les AIPs, et dans la mesure où elles ne représentent que des caractères alphanumériques, un codec simple basé sur une unique table de concordance soit employé. Une approche avec des codons de quatre nucléotides, telle que proposée par exemple par Agrawal et al. (2012), par Jiménez-Sánchez (2013) ou encore par Ubaidur Rahman et al. (2015), permet de faire correspondre l'entièreté de l'alphabet anglais, les dix

chiffres ainsi que d'autres symboles ; selon nous, la proposition d'A. Jiménez-Sánchez paraît la plus aboutie, mais pourra nécessiter quelques changements, voire son adaptation sur une base de cinq nucléotides pour éviter les quelques homopolymères qu'elle comporte. De cette manière, des métadonnées au format XML, telles que celles utilisées par la technologie DLCM, ou JSON peuvent être facilement encodées et décodées avec une unique page d'instruction, en contournant le problème de l'interprétation des données numériques. Pour chaque scénario, les modalités de différenciation des métadonnées et des données seront précisées.

Enfin, et lorsque les coûts de synthèse auront suffisamment chuté, il pourrait devenir viable de procéder périodiquement à une nouvelle synthèse de l'ensemble de l'archive afin d'éviter la multiplication des codecs au cours du temps en s'assurant qu'à chaque nouvelle synthèse tous les AIPs en utilisent la même version. Une nouvelle synthèse périodique pourrait également profiter d'éventuelles nouvelles avancées des biotechnologies, qui permettraient par exemple la synthèse de brins plus longs que 200 nucléotides, réduisant ainsi le nombre de chaînes de nucléotides par AIP. Enfin, une nouvelle synthèse permettrait également de mitiger les biais induits par une PCR. Cette approche différerait des remplacements périodiques des supports de stockage actuels en ce que contrairement à eux elle ne vient pas contrebalancer une faible longévité du support qui compromettrait l'intégrité même des données. Alors que les anciens brins restent tout à fait viables pour récupérer les données, une nouvelle synthèse permettrait d'éviter l'héritage d'anciennes pratiques dont la maintenance en parallèle des autres serait coûteuse.

Enfin, il nous paraît souhaitable d'inclure au sein de l'archive quelques brins d'ADN identifiables par une amorce particulière et dont le séquençage permettrait de valider que le codec à utiliser est bien le bon, particulièrement pour les codecs plus complexes encodant les bits.

5.4.5 Format des données

Comme la revue de la littérature dédiée (voir chapitre 4) a pu le laisser entrevoir, l'interprétation correcte des suites de bits afin de lire le fichier représente un enjeu majeur de la préservation à long terme des données numériques. À ce titre, il paraît indispensable de fournir la documentation adéquate pour continuer à lire ces formats : pour cela, le projet DNAMIC se base sur le maintien à disposition de la base PRONOM (Burgi 2024, p. 13). Toutefois, comme reconnu par le projet lui-même, cette approche n'est pas sans problèmes, au premier lieu la pérennité de cette base.

Pour répondre au problème de la préservation à long terme des données numériques, il nous paraît nécessaire que l'archive adopte des politiques en ce sens. D'une part, il est indispensable que l'archive réduise autant que possible le nombre de formats qu'elle est amenée à prendre en charge, en favorisant des formats libres, ouverts et éprouvés. Un petit nombre de formats réduira le risque qu'un format rare ne puisse plus être lisible, de même que cela réduira la documentation nécessaire. Le choix de formats libres et ouverts garantira que leur documentation puisse être incluse sans problèmes liés aux droits d'auteurs, tandis que le choix de formats éprouvés pourra peut-être permettre à la connaissance de ceux-ci de perdurer malgré la perte de la documentation au sein de l'archive.

D'autre part, il est nécessaire que l'archive destinant ses documents à la préservation à long terme implémente et archive des émulateurs permettant de relire ces documents, ne serait-ce que pour atteindre une réelle autonomie en éliminant toute dépendance externe pour accéder

aux données. À ce titre, limiter la diversité des formats réduira le nombre d'émulateurs à implémenter. Une approche telle que la machine virtuelle Chifir (Nguyen, Kay 2015), privilégiant la simplicité d'implémentation à son efficacité, déclinée en autant de formats que nécessaires, nous semble préférable à une unique machine virtuelle dont l'implémentation serait compliquée.

5.5 Spécificités pour le scénario simplifié

Pour rappel, le scénario simplifié prévoit d'implémenter un plan de reprise d'activité en tenant en compte les paramètres suivants : toute la stratégie de stockage passe par des séquences de nucléotides, sans recours aux nanostructures. De plus, les AIPs restent complets et ne sont pas divisés selon leurs différents éléments ; ils sont empaquetés dans un fichier ZIP. Optionnellement, certaines métadonnées peuvent être encodées et stockées physiquement à part (Burgi 2024, pp. 14-15).

5.5.1 Transmission de l'amorce générique et du codec

Dans le cas du scénario simplifié, toutes les données sont encodées sur des séquences de nucléotides. L'unique moyen d'accéder aux documents est donc de séquencer les brins d'ADN (Bornholt et al. 2016), puis de décoder les nucléotides pour obtenir les bits. Les techniques de séquençage ayant été couvertes par la documentation, il reste à transmettre l'amorce générale nécessaire à l'amplification de toute l'archive ainsi que le codec. Pour ce faire, le projet DNAMIC envisage que le tube à essai contenant l'archive soit libellé avec une indication sur le codec et sur l'amorce générique ; dans l'hypothèse où les libellés deviendraient illisibles, une autre méthode serait d'avoir recours à un secteur zéro permettant de retrouver cette information (Burgi 2024, p. 17).

Compte tenu de la place disponible sur un tube à essai, il serait sans doute possible d'y indiquer l'amorce générique, d'environ vingt nucléotides, mais la place manquerait sans doute pour y spécifier l'entièreté du codec : le libellé ne peut donc que pointer vers une information qui serait contenue dans un autre document. Nous avons justement indiqué précédemment que les informations relatives au codec et à la structure du brin devaient faire partie de la documentation sauvegardée en format analogique, vers laquelle le libellé du tube à essai peut pointer. Dans l'hypothèse où plusieurs codecs sont utilisés, une telle approche pourrait également convenir à un nombre restreint de codecs.

Toutefois, dans l'hypothèse où les libellés du tube à essai deviendraient illisibles, il faut avoir recours à d'autres stratégies. Plusieurs approches complémentaires peuvent s'entrevoir. D'une part, la documentation devrait maintenir au format analogique une table des contenants, des amorces génériques et des codecs, afin de savoir où est utilisé quoi. Idéalement, le tube à essai pourrait être gravé à sa surface avec un identifiant permettant de le distinguer de manière stable et non-ambiguë : si l'amorce et le codec indiqués sur le libellé pourraient changer au cours de la durée de vie de l'archive, l'identifiant du tube à essai devrait rester fixe.

D'autre part et comme prévu par le projet DNAMIC en conformité avec les standards de la DNA Data Storage Alliance, un brin d'ADN contenu au sein de l'archive pourrait permettre de faire le renvoi. Comme exprimé précédemment toutefois, nous ne pensons pas que ce renvoi doive être fait via une table de concordance maintenue sur Internet : le renvoi devrait plutôt être vers la documentation associée à l'archive. La transmission de l'amorce du secteur zéro dans la documentation permettrait d'amplifier le brin, même en l'absence de l'amorce générique.

Toutefois, nous souhaitons proposer une autre méthode, basée sur l'inclusion dans l'archive ADN de métadonnées de gestion, que nous entendons comme des métadonnées propres à l'archive en tant que telle et non aux documents archivés. Ces métadonnées, encodées par un codec simple basé sur une unique table de concordance, tel que présenté dans la section 5.4.4, pourraient contenir les spécifications du ou des codecs, de même que toute autre documentation nécessaire, y compris celle déjà fournie sur support analogique. Ces métadonnées de gestion devraient être accessibles via une amorce générique distincte de celle utilisée pour les documents : on créerait ainsi des espaces de nom (*namespace*). Nous entendons un espace de nom comme un cadre qui permet de regrouper et d'organiser des identificateurs sous une étiquette commune. Il permet de distinguer des éléments appartenant à des catégories ou à des contextes différents, facilitant ainsi leur gestion et leur accès au sein d'un ensemble logique unifié. Une amorce générique distincte en fonction du type d'information que transmet l'oligonucléotide permettrait la coexistence de données de différentes natures au sein du même ensemble logique qu'est la solution contenue dans le tube à essai.

Le libellé principal du tube à essai pourrait ainsi se limiter à indiquer l'amorce générique des métadonnées de gestion et le renvoi à la table de concordance du codec simple dans la documentation analogique, voire à imprimer la table de concordance complète si sa taille est suffisamment réduite. Ces métadonnées de gestion pourraient servir de secteur zéro en étant amplifiée en premier et en servant de base à l'accès au reste des documents. Toutefois, cette approche ne résout pas la nécessité de savoir quels brins doivent être décodés avec quel codec.

5.5.2 Gestion des amorces

Les diverses propositions que nous avons faites précédemment demandent une gestion fine des amorces, telle que présentée par Tomek et al. (2019) et par Song et al. (2021). Ces deux études se basent sur une PCR emboîtée (*nested PCR*) ou semi-emboîtée (*semi-nested PCR*), consistant à utiliser plusieurs paires d'amorces les unes après les autres au cours de différentes PCRs (Green, Sambrook 2019). Cette méthode a originellement été développée pour atteindre une meilleure précision et éliminer des séquences non voulues ; la PCR semi-emboîtée se distingue par la réutilisation d'une des amorces : dans le cadre de ce travail, nous ne ferons pas de distinction entre la PCR emboîtée et semi-emboîtée.

Dans le cadre du stockage sur ADN, la PCR emboîtée permet de résoudre deux problèmes majeurs. D'une part, selon Organick et al. (2018, Supplementary Note 5; voir aussi Tomek et al. 2019, fig. 1) il n'est possible de créer que 28 000 amorces de vingt bases suffisamment distinctes pour éviter des interactions indésirables entre elles ; en revanche, l'utilisation séquentielle de deux amorces permet d'atteindre près de deux cents millions de combinaisons possibles. D'autre part, une conception basée sur la combinaison d'amorces pourrait permettre à certaines d'entre elles de représenter des métadonnées basiques, facilitant ainsi la récupération des documents en fonction de ces informations. Par exemple, certaines amorces pourraient indiquer si l'oligonucléotide contient uniquement des métadonnées ou le contenu complet de l'AIP, ou encore spécifier le codec utilisé.

Naturellement, la combinaison de plusieurs amorces sur un même brin réduit d'autant la place restante pour les données elles-mêmes et donc la densité d'information. Une amorce de longueur comprise entre dix-huit et vingt-quatre base est généralement recommandée pour toute utilisation (Dieffenbach, Lowe, Dveksler 1993). Organick et al. (2018, Supplementary

Note 2) rapportent l'utilisation d'amorces de longueur comprise entre vingt et trente-trois bases pour les études portant sur le stockage sur ADN. Sur un brin de deux cents bases, plus d'un cinquième est donc fréquemment réservé pour encoder les amorces. Les études de Tomek et al. (2019) et de Song et al. (2021) utilisent toutes deux des amorces de vingt bases pour la PCR emboîtée : en y rajoutant l'amorce générique, on atteint donc 120 bases utilisées pour les amorces, soit les trois cinquièmes d'un oligonucléotide.

Toutefois, il faut rappeler que la limite de deux cents bases par brin est avant tout technique et que les technologies de synthèse se développant, la proportion prise par les amorces sur un oligonucléotide ira nécessairement en décroissant ; Yazdi et al. (2017) rapportent d'ailleurs avoir déjà utilisé des brins d'une longueur de 1000 nucléotides. De plus, certaines conceptions d'amorces pourraient permettre d'en réduire l'ampleur. Dieffenbach, Lowe et Dveksler (1993) précisent que pour de l'ADN non génomique, l'amorce peut être inférieure à quinze bases. D'autres chercheurs ont montré qu'une PCR peut fonctionner avec un minimum de sept bases (Vincent, Gurling, Melmer 1994). Une amorce trop petite est cependant au risque que sa séquence puisse se retrouver ailleurs sur le brin : lors de la PCR, des interactions non désirées auront lieu et des segments d'ADN correspondant aux amorces mais situés à l'intérieur des oligonucléotides seront dupliqués. La probabilité qu'une séquence de taille n puisse se retrouver sur un brin de deux cents nucléotides est définie par l'équation suivante :

$$\frac{(200 - n) + 1}{4^n}$$

Si avec neuf nucléotides la probabilité chute déjà sous les 0.01%, en pratique il faut encore la multiplier par le nombre d'oligonucléotides qui formeront l'AIP. Sur la base des fichiers encodés par Organick et al. (2018), où le plus gros fichier est représenté par plus de trois millions de brins d'ADN, une probabilité similaire n'est atteinte qu'avec une amorce d'une vingtaine de nucléotides.

Si l'on souhaite utiliser des amorces de petite taille, il est donc nécessaire d'avoir une approche différente. Deux techniques pourraient être envisagées : le recours aux homopolymères ou l'usage de trois bases au lieu de quatre. Dans les deux cas, il s'agit de créer des amorces telles qu'elles ne pourront se retrouver ailleurs dans les séquences. Alternativement, ces amorces de petites tailles pourraient n'être que des « pseudo-amorces » et n'être destinées à être utilisées qu'en combinaison avec d'autres amorces plus complètes : leur taille n'aurait ainsi pas d'incidence.

On pourrait donc imaginer une concaténation d'amorce en trois ou quatre niveaux :

1. Un premier niveau est utilisé pour les amorces génériques définissant des espaces de nom, c'est-à-dire pour spécifier si le brin d'ADN encode des documents ou des métadonnées de gestion ; ce niveau d'amorce devrait être composé d'une vingtaine de nucléotides.
2. Un deuxième niveau, composé de « pseudo-amorces » plus courtes, transmet successivement le codec et l'information sur le contenu porté par les oligonucléotides en les concaténant (nous anticipons ici sur la section 5.5.3.2 ci-après). Si l'on imagine deux « pseudo-amorces » de cinq nucléotides, dix nucléotides pourraient ainsi suffire pour transmettre un nombre raisonnable de codecs et de type d'information. Ces amorces de deuxième niveau ne sont pas prévues pour être utilisées de manière indépendante, mais en combinaison avec les amorces d'un autre niveau.

3. Un troisième niveau transmet l'amorce spécifique de l'AIP. De nouveau, ce niveau d'amorce devrait être composé d'une vingtaine de nucléotides afin de bien différencier chaque AIP.
4. Si l'archive possède plus de 28 000 AIPs, on peut envisager de rajouter un quatrième niveau permettant un nombre exponentiellement plus élevé de possibilités d'adressage.

5.5.3 Gestion des métadonnées

Si un AIP est un fichier ZIP contenant à la fois l'Information de contenu et les métadonnées associées, l'accès à ces dernières n'est possible qu'en ouvrant le fichier ZIP, ce qui n'est faisable qu'après avoir séquencé l'ensemble des brins d'un AIP. Une proposition du projet DNAMIC est donc de séparer les métadonnées du contenu (Burgi 2024, p. 12). Cette approche permettrait de relire uniquement les métadonnées et, en cas de reprise d'activité, de reconstruire la base de données sans manipuler les documents eux-mêmes. Si les métadonnées comprennent également l'information d'identification de l'AIP et que l'algorithme la transformant en amorce est également transmis, il serait possible ensuite de reconstruire l'amorce de n'importe quel AIP sans la connaître a priori. Toutefois, d'un point de vue OAIS et dans une optique d'auto-description, une telle proposition ne peut à notre sens s'envisager que si la séparation est en réalité une copie : les métadonnées sont extraites et encodées à part, les AIPs originaux restent complets et contiennent également les métadonnées en plus du contenu.

5.5.3.1 Option 1 : stockage physiquement à part des métadonnées

Une des options du scénario simplifié envisagé par le projet DNAMIC et de répercuter cette séparation également au niveau physique, en stockant les métadonnées dans une fiole différente (Burgi 2024, p. 12). Toutefois, la multiplication des supports de stockage induit nécessairement une perte de densité d'information (Newman et al. 2019, p. 2). Dans une telle approche, les libellés des tubes à essai comprendraient également le type d'information archivé, données ou métadonnées. Dans la mesure où tout ce qui a trait au même document serait adressable avec la même amorce, il serait facile de faire des équivalences entre les tubes à essai.

Des approches similaires ont été proposées dans la littérature. Ainsi, la séparation de l'archive en plusieurs conteneurs a été utilisée par Bornholt et al. (2016), mais cela concernait également les données. De leur côté, Marinelli et al (2023, p. 7) projettent de stocker les métadonnées physiquement à part et sur un format analogique, ne conservant sur ADN que les données elles-mêmes. Dans une logique d'auto-description toutefois, cette approche ne paraît pas complètement viable.

5.5.3.2 Option 2 : stockage au sein du même conteneur avec duplication des métadonnées

Pour éviter une perte de densité, stocker les métadonnées avec les données semble préférable. Dans ce cas de figure, opter pour une concaténation des amorces permettant de distinguer le type d'information contenu dans le brin semble être une approche prometteuse. Ainsi, tout AIP conserverait ses métadonnées au sein du fichier ZIP, mais celles-ci seraient également dupliquées et encodées seules selon le codec simplifié. Les brins de l'AIP et des métadonnées seraient adressables par la même amorce de troisième niveau, mais différeraient par les amorces de deuxième niveau. Avec les spécifications présentées précédemment, les métadonnées seraient adressables de deux manières :

- La combinaison de l'amorce générique des documents, de l'amorce du codec simplifié et de l'amorce pour les métadonnées permet de récupérer toutes les métadonnées contenues dans l'archive.
- La combinaison de l'amorce pour les métadonnées et de l'amorce spécifique de l'AIP permet de récupérer les métadonnées d'un seul AIP.

5.5.3.3 Option 3 : stockage au sein d'un même conteneur sans duplication des métadonnées

L'option précédente induit un certain nombre de répétition, réduisant la densité d'information. Une troisième option consisterait à encoder les métadonnées de telle manière qu'elles ne soient plus dans le fichier ZIP, mais qu'elles y restent associées en encodant le tout en utilisant la même amorce de troisième niveau. Une telle option n'est que partiellement compatible avec les spécificités du scénario simplifié, toutefois nous pensons qu'en conservant une amorce commune une certaine forme d'encapsulation est maintenue. Cette option s'inspire librement du système proposé par Li et al. (2021), où un certain nombre de métadonnées sont présentes dans une en-tête précédent les données elles-mêmes : autant que nous puissions en juger toutefois, dans cette étude les métadonnées sont encodées avec le même codec que les données elles-mêmes.

Au bout du compte, cette option se distingue de la précédente surtout par l'élimination des métadonnées au sein de l'AIP. Par rapport au système de Li et al. (2021), cette option encode un même fichier avec deux codecs différents, mais dont les différentes parties (données et métadonnées) sont adressables de manière distincte par la concaténation des amorces.

5.5.4 Gestion de l'émulation

Avec les approches proposées précédemment, et en se basant sur la solution M-1CI proposée par Li et al. (2021), une solution pour archiver les logiciels et leurs émulateurs peut s'entrevoir par la création d'un espace de nom qui leur serait dédié. D'une part, il pourrait être possible d'encoder le logiciel et le bootstrap comme des données (Cochrane, Chadash 2022, p. 227) et la documentation associée comme une métadonnée avec un codec simplifié, le tout partageant une même amorce identificatrice. D'autre part, le lien entre les fichiers et le logiciel associé peut être exprimé dans les métadonnées des fichiers, qui pourraient par exemple comporter une information sur l'outil et sur l'amorce nécessaire pour l'atteindre.

Alternativement, en suivant l'option 1 pour les métadonnées, les logiciels et les émulateurs pourraient être encodés sur ADN et conservés physiquement à part.

5.5.5 Résumé

Le scénario simplifié se caractérise par l'usage exclusif des séquences de nucléotides pour le stockage des données et l'unicité de l'AIP. Pour répondre à ces contraintes, les recommandations que nous proposons s'appuient sur les amorces et les possibilités de les concaténer pour créer une forme rudimentaire d'indexation.

- Indication du codec et de l'amorce générique sur le libellé du tube à essai ainsi que dans la documentation, création d'un identifiant unique et pérenne pour les tubes à essai qui puisse servir de concordance en cas d'effacement du label.
- Création de trois espaces de noms par l'implémentation de trois amorces génériques, permettant la distinction entre les données et métadonnées liées au contenu, les métadonnées de gestion, pouvant transmettre des

paramètres et de la documentation sur l'archive, y compris le codec et les amorces génériques, et les logiciels et émulateurs nécessaires à l'interprétation des données une fois décodées.

- Concaténation des amorces jusqu'à quatre niveaux, incluant une amorce générique, des « pseudo-amorces » pour le codec et le type de fichier (non destinées à être utilisées seules, mais en combinaison avec le premier ou le troisième niveau) et une amorce pour l'identifiant de l'AIP. Les « pseudo-amorces » permettent de transmettre des indications sur la manière de décoder le brin ainsi que sur son contenu. Un quatrième niveau n'est présent que si le nombre d'amorces de troisième niveau est insuffisant par rapport au nombre d'AIPs.
- Choix d'un codec simplifié, basé sur une table de concordance, pour encoder les métadonnées alphanumériques type XML, éliminant les problèmes liés à l'interprétation des données binaires.
- Proposition de trois options pour la gestion des métadonnées, avec
 - stockage physiquement séparé des métadonnées
 - stockage au sein du même conteneur avec duplication des métadonnées
 - stockage au sein du même conteneur sans duplication des métadonnées
- Inclusion des logiciels et des émulateurs encodés sur ADN, avec un espace de nom distinct

5.6 Spécificités pour le scénario avancé

Pour rappel, le scénario avancé fait appel tant aux séquences de nucléotides qu'aux nanostructures. Les AIPs peuvent être transformés, notamment pour permettre un encodage fichier par fichier ou pour distinguer les métadonnées des données ; il n'y a pas d'obligation de conserver le tout dans un fichier ZIP. Les métadonnées doivent être impérativement encodées et stockées physiquement avec les données.

L'avantage principal du scénario avancé sur le scénario simplifié est l'usage des nanostructures, qui pourraient permettre d'accéder à l'information par d'autres procédés que l'amplification par PCR puis le séquençage, une méthode particulièrement lente et destructrice. Ces nanostructures peuvent être de types très variés (Zhan et al. 2023), mais nous pensons que pour ne pas perdre la caractéristique d'absence d'obsolescence liée à l'usage de l'ADN, ces nanostructures et les procédés permettant d'accéder à l'information qu'elles contiennent ne doivent pas être trop compliqués ou faire usage d'équipements particuliers. Ainsi, nous pensons qu'un choix de nanostructures dont les informations sont accessibles par des procédés optiques est plus durable que, par exemple, des solutions électrochimiques telle que celle proposée par Jimenez-Munoz et al. (2024).

Une des applications les plus prometteuses des nanostructures nous semble être l'usage de QR-codes à l'échelle du micromètre (Choi et al. 2020; Dickinson et al. 2021). Stockable dans la même solution que les oligonucléotides, ils permettent la lecture de leur contenu par des procédés optiques bien plus rapides que la PCR. Naturellement, il serait nécessaire d'inclure dans la documentation analogique toute la documentation nécessaire à la lecture des QR-codes. Alternativement, l'usage de nanostructures pour créer des formes particulières permettrait également de transmettre un plus petit nombre d'information. Dans les deux cas,

la masse des données à archiver sera sans doute toujours conservée dans des chaînes de nucléotides.

Dans cette section, nous explorerons l'impact que les spécificités du scénario avancé ont par rapport au scénario simplifié.

5.6.1 Option 1 : utilisation de QR-codes

Au moins deux études ont utilisé des QR-codes ou s'en sont inspirées pour créer des nanostructures ADN. Choi et al. (2020) ont proposé une solution où les QR-codes ne contiennent que des métadonnées, mais où les oligonucléotides constituant les données sont physiquement rattachés aux QR-codes, permettant une forme d'autonomie des différents fichiers au sein même de l'archive. Dickinson et al. (2021) ont quant à eux proposé un concept où les QR-codes contiennent toute l'information. Si cette seconde étude offre l'avantage de complètement se passer du séquençage, sa densité de stockage d'information apparaît bien moindre que la première, qui nous semble plus intéressante et sur laquelle nous nous basons.

Il convient de noter que, dans leurs spécifications actuelles, les QR-codes sont eux-mêmes limités dans la quantité d'information qu'ils peuvent contenir. Les plus gros modules (version 40, 177x177) ne peuvent stocker qu'un maximum de 4296 caractères alphanumériques ou 2953 bytes, ceci avec une correction d'erreur limitée. Au sein de la technologie DLCM, ces métadonnées sont actuellement encodées en XML : les caractères alphanumériques ne comportant par les chevrons nécessaires aux balises XML, il serait donc nécessaire d'avoir recours aux bytes. Ainsi, selon l'ampleur des métadonnées d'un AIP, il sera sans doute nécessaire d'avoir recours à plusieurs QR-codes pour transmettre l'intégralité des métadonnées. Ceci ne semble pas être nécessairement un problème, l'étude de Choi et al. (2020) montrant qu'il est possible de les accoler les uns aux autres.

Dans une telle approche, les métadonnées accessibles par QR-code contiendraient à la fois les métadonnées de gestion, dont nous avons parlé précédemment dans le cadre du scénario simplifié, les métadonnées des AIPs et les métadonnées des outils nécessaires à l'émulation ; dans chacun des cas, l'appartenance des métadonnées à l'un ou l'autre espace de nom devra être spécifiée dans les métadonnées elles-mêmes, voire le QR-code pourrait comporter une bordure stylisée afin de permettre la reconnaissance immédiate de l'espace de nom. Pratiquement, il serait ainsi suffisant de transmettre le symbole rattaché au QR-code de l'espace de nom dédié aux métadonnées de gestion pour permettre leur lecture et la livraison de toutes la documentation nécessaire à l'accès des données archivistiques. Ainsi, la documentation expliquant l'ADN et les techniques de séquençage, de même que les paramètres d'encodage et le codec des oligonucléotides n'auraient théoriquement besoin de n'être transmis que sous forme de QR-code. Toutefois, dans une optique de redondance, il paraît plus pérenne de conserver la documentation analogique.

Vis-à-vis des fichiers, les métadonnées comprendront l'amorce nécessaire à leur amplification. Ainsi, les oligonucléotides ne seront composés que d'amorces à deux niveaux, comprenant l'amorce générique de l'espace de nom concerné et l'amorce identificatrice de l'AIP.

5.6.2 Option 2 : utilisation de nanostructures avec des formes distinctes

Cette option tire partie des formes variées qu'il est possible de donner aux oligonucléotides, soit par les origamis ADN soit par la topologie, pour transmettre quelques métadonnées. Nous pensons qu'une telle approche pourrait être utilisée avec succès pour transmettre le codec,

les types de données ou les espaces de noms dont nous avons parlé dans la partie précédente : s'agissant d'informations importantes mais peu variées dans l'ensemble de l'archive, leur nombre serait limité et il serait possible de maintenir dans la documentation analogique une table de concordance des formes utilisées. Dans une telle option, il serait nécessaire de conserver les métadonnées de gestion, celles des AIPs et celles des outils d'émulation dans des oligonucléotides, qu'il serait nécessaire de séquencer pour lire. À certains égards, une telle option est en réalité très semblable à l'option trois du scénario simplifié : seules les « pseudo-amorces » sont éliminées au profit de nanostructures.

6. Conclusion

L'ADN en tant que support de stockage des données numériques représente une alternative prometteuse aux supports optiques ou magnétiques actuellement utilisés. Cette technologie pourrait offrir une solution durable face aux défis du stockage des données numériques : en effet, les propriétés de l'ADN, telles que sa densité de stockage élevée, sa stabilité à long terme et l'absence d'obsolescence en font un candidat sérieux pour le stockage des données sur des échelles de temps de plusieurs siècles voire millénaires. C'est pourquoi le projet DNAMIC est intéressé à développer un connecteur ADN pour permettre, au sein de la technologie DLCM, l'archivage à long terme des données de la recherche et des publications.

Toutefois, dans la perspective de la préservation à long terme des données dans un contexte archivistique, et plus particulièrement en appliquant le modèle OAIS pour lequel il est important d'avoir une maîtrise complète des données jusqu'au niveau des bits, le choix du support n'est qu'une première étape. En effet, afin de garantir que les données stockées sur ADN restent accessibles et utilisables par les générations futures, il est essentiel de développer un plan de reprise d'activité envisageant une interruption sur une durée indéterminée, qui serait suffisamment longue ou grave pour avoir entraîné une perte des connaissances nécessaires à l'accès aux données. Un tel plan de reprise d'activité passe nécessairement par des stratégies permettant d'assurer une autodescription des données ainsi que leur autonomie, en limitant autant que possible le recours à des dépendances externes.

Dans le cadre de ce mémoire, nous avons développé des recommandations pour le développement de ce connecteur ADN, selon deux scénarios aux caractéristiques différentes tant dans les formes que prend le stockage ADN que sur l'intégrité structurelle des AIPs. Un certain nombre de recommandations transcendent ces différences. D'une part, un archivage à long terme sur support ADN ne saurait se passer entièrement d'une documentation immédiatement perceptible par un être humain : des choix doivent être faits tant dans la matérialité de cette documentation (type de support) que dans les caractéristiques de sa rédaction (langue, technicité) et des documents à sélectionner. D'autre part, une réflexion doit être menée sur le choix du codec et sur les formats de données utilisés par l'archive, afin de privilégier des approches pouvant être documentées facilement. Plus généralement, nous pensons que tout projet d'archivage ADN doit avoir pour principe la simplicité et la redondance, afin de garantir sa pérennité.

Au sein du scénario simplifié, nous nous sommes intéressés à la transmission de l'amorce générique et du ou des codecs. Nous avons proposé une solution basée sur la concaténation des amorces et sur l'existence de plusieurs espaces de nom afin de créer une forme rudimentaire d'indexation et transmettre certaines informations importantes dès la lecture des nucléotides. Cette approche a également été étendue à toutes les métadonnées accessibles hors d'un AIP qui profitent ainsi d'un codec simplifié permettant la transmission de l'information sans passer par le binaire.

Au sein du scénario avancé, nous avons retenu les possibilités offertes par des nanostructures reproduisant des QR-codes et la possibilité d'utiliser des formes particulières s'appuyant sur des origamis ADN ou sur la topologie pour transmettre le codec et/ou le type d'information contenu dans l'AIP, ce afin d'offrir une plus grande rapidité d'accès à l'information et une meilleure densité par rapport au scénario simplifié.

Nous reconnaissons que les recommandations que nous avons formulées pour les deux scénarios restent théoriques. Il sera nécessaire de les expérimenter en pratique afin de déterminer si leur implémentation est souhaitable ou si elles soulèvent des questions qui n'avaient pas été prévues. À ce titre, il est essentiel de poursuivre les efforts de recherche et de développement dans ce domaine dans une optique multidisciplinaire, en associant des archivistes aux informaticiens et aux biologistes afin que des décennies d'expertise dans la préservation numérique puisse être dès maintenant appliquées au stockage sur ADN.

Bibliographie

AGRAWAL, Akanksha et al., 2012. Implementation of DNA algorithm for secure voice communication. *International Journal of Scientific & Engineering Research*. Vol. 3, no 6, pp. 1140-1144.

AILENBERG, Menachem et ROTSTEIN, Ori D., 2009. An improved Huffman coding method for archiving text, images, and music characters in DNA. *BioTechniques*. Vol. 47, no 3, pp. 747-754. DOI 10.2144/000113218.

ANAVY, Leon et al., 2019. Data storage in DNA with fewer synthesis cycles using composite DNA letters. *Nature Biotechnology*. Vol. 37, no 10, pp. 1229-1236. DOI 10.1038/s41587-019-0240-x.

ANDRIAMAHADY, Dina, 2021. *OAIS compliant digital archiving in DNA* [en ligne]. Genève : Haute école de gestion de Genève. Disponible à l'adresse : <https://sonar.ch/global/documents/314970> [consulté le 15 août 2024].

ANŽEL, Aleksandar, HEIDER, Dominik et HATTAB, Georges, 2021. The visual story of data storage: From storage properties to user interfaces. *Computational and Structural Biotechnology Journal*. Vol. 19, pp. 4904-4918. DOI 10.1016/j.csbj.2021.08.031.

APPUSWAMY, Raja et JOGUIN, Vincent, 2021. Universal layout emulation for long-term database archival. In : *11th Annual Conference on Innovative Data Systems Research* [en ligne]. Chaminade. 2021. Disponible à l'adresse : https://www.cidrdb.org/cidr2021/papers/cidr2021_paper30.pdf [consulté le 11 août 2024].

BANCROFT, Carter et al., 2001. Long-Term Storage of Information in DNA. *Science*. Vol. 293, no 5536, pp. 1763-1765. DOI 10.1126/science.293.5536.1763c.

BEARMAN, David, 1999. Reality and Chimeras in the Preservation of Electronic Records. *D-Lib Magazine* [en ligne]. Vol. 5, no 4. Disponible à l'adresse : <http://www.dlib.org/dlib/april99/bearman/04bearman.html> [consulté le 11 août 2024].

BIBLIOTHÈQUE NATIONALE DE FRANCE, 2024. Formats de données pour la préservation numérique. *BnF - Site institutionnel* [en ligne]. 2024. Disponible à l'adresse : <https://www.bnf.fr/fr/formats-de-donnees-pour-la-preservation-numerique> [consulté le 11 août 2024].

BORNHOLT, James et al., 2016. A DNA-Based Archival Storage System. In : *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, pp. 637-649. Atlanta Georgia USA : ACM. 25 mars 2016. ISBN 978-1-4503-4091-5. DOI 10.1145/2872362.2872397.

BURGI, Pierre Yves et MAKHLOUF SHABOU, Basma, 2021. Le projet Data Life-Cycle Management (DLCM) en Suisse : une gestion des données de la recherche pensée pour ses utilisateurs. *I2D - Information, données & documents*. Vol. 2, no 2, pp. 87-95. DOI 10.3917/i2d.212.0087.

BURGI, Pierre-Yves et al., 2022. OAIS-compliant digital archiving of research and patrimonial data in DNA. In : *Proceedings of the 18th International Conference on Digital Preservation*, pp. 220-224 [en ligne]. Glasgow. 2022. Disponible à l'adresse : <https://archive-ouverte.unige.ch/unige:165826> [consulté le 15 août 2024].

BURGI, Pierre-Yves, 2024. *2.1 Framework architecture specifications*. . Genève : Université de Genève. [document interne]

CALLA, Simon et al., 2023. Confronting the Uncertainties Associated with Long-Time Scales: Analysis of the Modes of Preservation of Memory of Radioactive Waste Burial Sites. *Worldwide Waste*. Vol. 6, no 1, pp. 1-12. DOI 10.5334/wwwj.75.

CAPOVA, Klara Anna, 2021. Introducing Humans to the Extraterrestrials: the Pioneering Missions of the Pioneer and Voyager Probes. *Frontiers in Human Dynamics*. Vol. 3, pp. 1-12. DOI 10.3389/fhumd.2021.714616.

CCSDS, 2019. *Reference Model for an Open Archival Information System (OAIS)*. . CCSDS 650.0-P-3. CCSDS 650.0-P-3.

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

CHOI, Yeongjae et al., 2019. High information capacity DNA-based data storage with augmented encoding characters using degenerate bases. *Scientific Reports*. Vol. 9, no 1, p. 6582. DOI 10.1038/s41598-019-43105-w.

CHOI, Yeongjae et al., 2020. DNA Micro-Disks for the Management of DNA-Based Data Storage with Index and Write-Once–Read-Many (WORM) Memory Features. *Advanced Materials*. Vol. 32, no 37, pp. 1-8. DOI 10.1002/adma.202001249.

CHURCH, George M., GAO, Yuan et KOSURI, Sriram, 2012. Next-Generation Digital Information Storage in DNA. *Science*. Vol. 337, no 6102, pp. 1628-1629. DOI 10.1126/science.1226355.

CLELLAND, Catherine Taylor, RISCA, Viviana et BANCROFT, Carter, 1999. Hiding messages in DNA microdots. *Nature*. Vol. 399, no 6736, pp. 533-534. DOI 10.1038/21092.

COCHRANE, Euan et CHADASH, Daniel, 2022. DNA Data Storage for Long Term Digital Preservation. In : *Proceedings of the 18th International Conference on Digital Preservation*, pp. 225-230 [en ligne]. Glasgow. 2022. Disponible à l'adresse : <https://phaidra.univie.ac.at/detail/o:1893668> [consulté le 15 août 2024].

COX, Jonathan P. L., 2001. Long-term data storage in DNA. *Trends in Biotechnology*. Vol. 19, no 7, pp. 247-250. DOI 10.1016/S0167-7799(01)01671-7.

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

DICKINSON, George D. et al., 2021. An alternative approach to nucleic acid memory. *Nature Communications*. Vol. 12, pp. 1-12. DOI 10.1038/s41467-021-22277-y.

DIEFFENBACH, C. W., LOWE, T. M. et DVEKSLER, G. S., 1993. General concepts for PCR primer design. *PCR Methods and Applications*. Vol. 3, no 3, pp. S30-S37.

DNAMIC, 2024. DNAMIC. [en ligne]. 2024. Disponible à l'adresse : <https://dnamic.org/> [consulté le 15 août 2024].

DORICCHI, Andrea et al., 2022. Emerging Approaches to DNA Data Storage: Challenges and Prospects. *ACS Nano*. Vol. 16, no 11, pp. 17552-17571. DOI 10.1021/acsnano.2c06748.

DUMONT, Jean-Noël et CHARTON, Patrick, 2012. ANDRA's long term memory experiences and programme – J-N. Dumont & P. Charton. In : *The Preservation of Records, Knowledge and Memory (RK&M) Across Generations: Scoping the Issue. Workshop Proceedings, Issy-les-Moulineaux, France, 11-13 October 2011*, pp. 28-29 [en ligne]. Paris : Nuclear Energy

Agency. 2012. Disponible à l'adresse : https://www.oecd-nea.org/jcms/pl_39202/the-preservation-of-records [consulté le 15 août 2024].

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

ETH ZURICH, 2024. File formats for archiving. *Research Data Management and Digital Curation Research Data Management and Digital Curation* [en ligne]. 19 juin 2024. Disponible à l'adresse : <https://unlimited.ethz.ch/display/DD/File+formats+for+archiving> [consulté le 15 août 2024].

EVANS, R. F. L. et al., 2012. Thermally induced error: Density limit for magnetic data storage. *Applied Physics Letters*. Vol. 100, no 10, pp. 1-4. DOI 10.1063/1.3691196.

EXTANCE, Andy, 2016. How DNA could store all the world's data. *Nature*. Vol. 537, no 7618, pp. 22-24. DOI 10.1038/537022a.

GARAFUTDINOV, Ravil R. et al., 2022. Encoding of non-biological information for its long-term storage in DNA. *Biosystems*. Vol. 215-216, p. 1è8. DOI 10.1016/j.biosystems.2022.104664.

GAVREL, Sue, 1986. Preserving Machine-Readable Archival Records: A Reply to John Mallinson. *Archivaria*. Vol. 22, pp. 153-155.

GOLDMAN, Nick et al., 2013. Towards practical, high-capacity, low-maintenance information storage in synthesized DNA. *Nature*. Vol. 494, no 7435, pp. 77-80. DOI 10.1038/nature11875.

GRASS, Robert N. et al., 2015. Robust Chemical Preservation of Digital Information on DNA in Silica with Error-Correcting Codes. *Angewandte Chemie International Edition*. Vol. 54, no 8, pp. 2552-2555. DOI 10.1002/anie.201411378.

GREEN, Michael R. et SAMBROOK, Joseph, 2019. Nested Polymerase Chain Reaction (PCR). *Cold Spring Harbor Protocols*. Vol. 2019, no 2, pp. 175-178. DOI 10.1101/pdb.prot095182.

HEMINGER, Alan R. et KELLEY, Don M., 2005. Assessing the Digital Rosetta Stone Model for Long-Term Access to Digital Documents. *Journal of Management Information Systems*. Vol. 21, no 4, pp. 11-35. DOI 10.1080/07421222.2005.11045827.

HEMINGER, Alan R. et ROBERTSON, Steven B., 2000. The Digital Rosetta Stone: A Model for Maintaining Long-term Access to Static Digital Documents. *Communications of the Association for Information Systems*. Vol. 3, no 1. DOI 10.17705/1CAIS.00302.

HILBERT, Martin et LÓPEZ, Priscila, 2011. The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*. Vol. 332, no 6025, pp. 60-65. DOI 10.1126/science.1200970.

HORMANN, Peter et CAMPBELL, Leith, 2014. Data Storage Energy Efficiency in the Zettabyte Era. *Australian Journal of Telecommunications and the Digital Economy*. Vol. 2. DOI 10.7790/ajtde.v2n3.51.

ISO, 1994. *Information et documentation - Papier pour documents - Prescriptions pour la permanence*. ISO 9706. International Organization for Standardization. International Organization for Standardization ISO 9706.

ISO, 2012. *Space data and information transfer systems — Open archival information system (OAIS) — Reference model*. ISO 14721. International Organization for Standardization. International Organization for Standardization ISO 14721.

JIANG, Jonathan H. et al., 2023. Message in a Bottle—An Update to the Golden Record: 1. Objectives and Key Content of the Message. *Earth and Space Science*. Vol. 10, no 12, pp. 1-11. DOI 10.1029/2023EA003042.

JIMENEZ-MUNOZ, Miguel A., WOOD, Christopher et WÄLTI, Christoph, 2024. Towards the development of an electrochemical random access DNA memory (e-RADM). *MRS Advances*. Vol. 9, no 8, pp. 531-536. DOI 10.1557/s43580-024-00784-6.

JIMÉNEZ-SÁNCHEZ, Alfonso, 2013. A proposal for a DNA-based computer code. *International Invention Journal of Biochemistry and Bioinformatics*. Vol. 1, pp. 1-4.

JOGUIN, Vincent, 2019. Passive Digital Preservation Now & Later: Microfilm, Micr'Olonys and DNA. In : RAS, Marcel, SIEMAN, Barbara et PUGGIONI, Angela (éd.), *Proceedings of the 16th International Conference on Digital Preservation*, pp. 354-359 [en ligne]. Amsterdam. 2019. ISBN 978-90-6259-043-8. Disponible à l'adresse : https://ipres2019.org/static/pdf/iPres2019_paper_139.pdf

JOGUIN, Vincent et DUMONT, Jean-Noël, 2022. Passive Digital Preservation on Paper in Practice. In : *Proceedings of the 18th International Conference on Digital Preservation*, pp. 271-276 [en ligne]. Glasgow. 2022. Disponible à l'adresse : <https://phaidra.univie.ac.at/detail/o:1893668> [consulté le 15 août 2024].

JONES, Nicola, 2018. How to stop data centres from gobbling up the world's electricity. *Nature*. Vol. 561, no 7722, pp. 163-166. DOI 10.1038/d41586-018-06610-y.

KELLY, Kevin, 2008. Very Long-Term Backup. *Long Now* [en ligne]. 20 août 2008. Disponible à l'adresse : <https://longnow.org/ideas/very-long-term-backup/> [consulté le 15 août 2024].

LABARCA, Joseph E., 2012. Preservation of Documents and Photographic Images: Long Term Strategies for Future Generations. In : *Archiving 2011 Final Program and Proceedings*, pp. 136-143 [en ligne]. Salt Lake City : Society for Imaging Science and Technology. 2012. ISBN 978-1-63266-640-6. Disponible à l'adresse : <https://library.imaging.org/admin/apis/public/api/ist/website/downloadArticle/archiving/8/1/art00032>

LAVOIE, Brian, 2014. *The Open Archival Information System (OAIS) Reference Model: Introductory Guide (2nd Edition)*. Digital Preservation Coalition. DOI 10.7207/twr14-02.

LI, Min et al., 2021. A self-contained and self-explanatory DNA storage system. *Scientific Reports*. Vol. 11, no 1, pp. 1-15. DOI 10.1038/s41598-021-97570-3.

LIBRARY OF CONGRESS, 2024. Recommended Formats Statement. *Library of Congress - Preservation* [en ligne]. 2024. Disponible à l'adresse : <https://www.loc.gov/preservation/resources/rfs/TOC.html> [consulté le 15 août 2024].

LORIE, Raymond A., 2001. Long term preservation of digital information. In : *Proceedings of the 1st ACM/IEEE-CS joint conference on Digital libraries*, pp. 346-352. New York, NY, USA : Association for Computing Machinery. juin 2001. JCDL '01. ISBN 978-1-58113-345-5. DOI 10.1145/379437.379726.

LORIE, Raymond A. et VAN DIESEN, Raymond J., 2005. *UVC: A Universal Virtual Computer for Long-Term Preservation of Digital Information* [en ligne]. IBM. RJ 10338. IBM

Research Report. Disponible à l'adresse :
<https://dominoweb.draco.res.ibm.com/reports/rj10338.pdf>

LUNT, Barry M, 2012. How Long is Long-Term Data Storage? In : *Archiving 2011 Final Program and Proceedings*, pp. 29-33 [en ligne]. Salt Lake City : Society for Imaging Science and Technology. 2012. ISBN 978-1-63266-640-6. Disponible à l'adresse :
https://www.imaging.org/common/uploaded%20files/pdfs/Reporter/Articles/2011_26/REP26_3_4_ARCH2011_Lunt.pdf

MALLINSON, John C., 1986. Preserving Machine-Readable Archival Records for the Millennia. *Archivaria*. No 22, pp. 147-152.

MARINELLI, Eugenio et al., 2023. Towards Migration-Free « Just-in-Case » Data Archival for Future Cloud Data Lakes Using Synthetic DNA. *Proceedings of the VLDB Endowment*. Vol. 16, no 8, pp. 1923-1929. DOI 10.14778/3594512.3594522.

MASANET, Eric et al., 2020. Recalibrating global data center energy-use estimates. *Science*. Vol. 367, no 6481, pp. 984-986. DOI 10.1126/science.aba3758.

NEIMAN, Mikhail Samoilovich, 1964. Некоторые принципиальные вопросы микроминиатюризации. *Радиотехника*. Vol. 19, no 1, pp. 3-12.

NEWMAN, Sharon et al., 2019. High density DNA data storage library via dehydration with digital microfluidic retrieval. *Nature Communications*. Vol. 10, no 1, pp. 1-6. DOI 10.1038/s41467-019-09517-y.

NGUYEN, Bichlien et al., 2020. Architecting Datacenters for Sustainability: Greener Data Storage using Synthetic DNA. In : *Electronics Goes Green 2020* [en ligne]. 1 septembre 2020. Disponible à l'adresse : <https://www.microsoft.com/en-us/research/publication/architecting-datacenters-for-sustainability-greener-data-storage-using-synthetic-dna/> [consulté le 15 août 2024].

NGUYEN, Long Tien et KAY, Alan, 2015. The cuneiform tablets of 2015. In : *2015 ACM International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software (Onward!)*, pp. 297-307. Pittsburgh PA USA : ACM. 21 octobre 2015. ISBN 978-1-4503-3688-8. DOI 10.1145/2814228.2814250.

NUCLEAR ENERGY AGENCY, 2019a. *Preservation of Records, Knowledge and Memory Across Generations: Final Report* [en ligne]. Paris : Nuclear Energy Agency. 7421. Disponible à l'adresse : https://www.oecd-neo.org/jcms/pl_15088 [consulté le 15 août 2024].

NUCLEAR ENERGY AGENCY, 2019b. *Preservation of Records, Knowledge and Memory Across Generations: Developing a Key Information File for a Radioactive Waste Repository* [en ligne]. Paris : Nuclear Energy Agency. 7377. Disponible à l'adresse : https://www.oecd-neo.org/jcms/pl_15060/preservation-of-records-knowledge-and-memory-across-generations-developing-a-key-information-file-for-a-radioactive-waste-repository?details=true [consulté le 15 août 2024].

NUCLEAR ENERGY AGENCY, 2019c. *Preservation of Records, Knowledge and Memory (RK&M) Across Generations: Compiling a Set of Essential Records for a Radioactive Waste Repository* [en ligne]. Paris : Nuclear Energy Agency. 7423. Disponible à l'adresse : https://www.oecd-neo.org/jcms/pl_15090/preservation-of-records-knowledge-and-memory-rk-m-across-generations-compiling-a-set-of-essential-records-for-a-radioactive-waste-repository?details=true [consulté le 15 août 2024].

O' DRISCOLL, Aisling et SLEATOR, Roy D., 2013. Synthetic DNA: The next generation of big data storage. *Bioengineered*. Vol. 4, no 3, pp. 123-125. DOI 10.4161/bioe.24296.

ORGANICK, Lee et al., 2018. Random access in large-scale DNA data storage. *Nature Biotechnology*. Vol. 36, no 3, pp. 242-248. DOI 10.1038/nbt.4079.

ORGANICK, Lee et al., 2020. Probing the physical limits of reliable DNA data retrieval. *Nature Communications*. Vol. 11, no 1, p. 616. DOI 10.1038/s41467-020-14319-8.

PANDA, Darshan et al., 2018. DNA as a digital information storage device: hope or hype? *3 Biotech*. Vol. 8, no 5, pp. 1-9. DOI 10.1007/s13205-018-1246-7.

POPHAM, Michael et MITCHAM, Jenny, 2022. From Ray cats to DPC RAM: How Best to Preserve a Digital Memory of the Nuclear Decommissioning Process. In : *Proceedings of the 18th International Conference on Digital Preservation*, pp. 315-318 [en ligne]. Glasgow. 2022. Disponible à l'adresse : <https://phaidra.univie.ac.at/detail/o:1893668> [consulté le 15 août 2024].

REINSEL, David, GANTZ, John et RYDNING, John, 2018. *The Digitization of the World from Edge to Core* [en ligne]. IDC. White paper . Disponible à l'adresse : <https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf> [consulté le 15 août 2024].

REN, Yubin et al., 2022. DNA-Based Concatenated Encoding System for High-Reliability and High-Density Data Storage. *Small Methods*. Vol. 6, no 4, pp. 1-9. DOI 10.1002/smt.202101335.

REY, Alain, 2013. The Nanoform : Sapphire Disk System of Records Preservation used by ANDRA. In : *The Preservation of Records, Knowledge and Memory (RK&M) Across Generations: Scoping the Issue. Workshop Proceedings, Issy-les-Moulineaux, France, 11-13 October 2011*, pp. 81-85 [en ligne]. Paris : Nuclear Energy Agency. 2013. Disponible à l'adresse : https://www.oecd-neo.org/jcms/pl_19288/the-preservation-of-records-knowledge-and-memory-rk-m-across-generations-improving-our-understanding-workshop-proceedings [consulté le 15 août 2024].

ROSENTHAL, David S. H. et al., 2013. The Economics of Long-Term Digital Storage. In : DURANTI, Luciana et SCHAFFER, Elizabeth (éd.), *The Memory of the World in the Digital Age: Digitization and Preservation. An International Conference on Permanent Access to Digital Documentary Heritage*, pp. 513-528 [en ligne]. Vancouver : UNESCO. 2013. Disponible à l'adresse : <https://unesdoc.unesco.org/ark:/48223/pf0000373728> [consulté le 15 août 2024].

ROSENTHAL, David Stuart Holmes, 2017. The medium-term prospects for long-term storage systems. *Library Hi Tech*. Vol. 35, no 1, pp. 11-31. DOI 10.1108/LHT-11-2016-0128.

ROTHENBERG, Jeff, 1995. Ensuring the Longevity of Digital Documents. *Scientific American*. Vol. 272, no 1, pp. 42-47.

ROTHENBERG, Jeff, 1998. Ensuring the Longevity of Digital Information. *International Journal of Legal Information*. Vol. 26, no 1-3, pp. 1-22. DOI 10.1017/S073112650000469.

ROTHENBERG, Jeff, 1999. *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation* [en ligne]. Washington, DC : Council on Library and Information Resources. ISBN 1-887334-63-7. Disponible à l'adresse : <https://www.clir.org/pubs/reports/rothenberg/> [consulté le 20 mai 2024].

RUMMELHOFF, Ivar et al., 2021. An Abstract Machine Approach to Preserving Digital Information. *IEEE Access*. Vol. 9, pp. 154914-154932. DOI 10.1109/ACCESS.2021.3128382.

- RUTTEN, Martin G. T. A. et al., 2018. Encoding information into polymers. *Nature Reviews Chemistry*. Vol. 2, no 11, pp. 365-381. DOI 10.1038/s41570-018-0051-5.
- SABLIŃSKI, Jędrzej et TRUJILLO, Alfredo, 2021. Piql. Long-term preservation technology study. *Archeion*. Vol. 122, pp. 13-32. DOI 10.4467/26581264ARC.21.011.14491.
- SAGAN, Carl, SAGAN, Linda Salzman et DRAKE, Frank, 1972. A Message from Earth. *Science*. Vol. 175, no 4024, pp. 881-884. DOI 10.1126/science.175.4024.881.
- SITAKER, Kragen Javier, 2018. Bootstrapping instruction set. [en ligne]. 6 novembre 2018. Disponible à l'adresse : <https://dercuano.github.io/notes/bootstrapping-instruction-set.html> [consulté le 15 août 2024].
- SMITH, Geoff C. et al., 2003. Some possible codes for encrypting data in DNA. *Biotechnology Letters*. Vol. 25, no 14, pp. 1125-1130. DOI 10.1023/A:1024539608706.
- SNIA, 2023a. *DNA Data Storage Sector Zero*. SNIA. SNIA. SNIA . Disponible à l'adresse : <https://www.snia.org/sites/default/files/technical-work/dna/release/DNA%20Data%20Storage%20Sector%20Zero%20v1.0.pdf> [consulté le 15 août 2024].
- SNIA, 2023b. *DNA Data Storage Sector One*. SNIA. SNIA. SNIA . Disponible à l'adresse : <https://www.snia.org/sites/default/files/technical-work/dna/release/DNA%20Data%20Storage%20Sector%20Zero%20v1.0.pdf> [consulté le 15 août 2024].
- SOLOMON, Richard Jay et al., 2021. WOLF (Write Once, Read Forever) Next Generation Archival Big Data Storage. In : *2021 IEEE Aerospace Conference*, pp. 1-7. Big Sky, MT, USA : IEEE. 6 mars 2021. ISBN 978-1-72817-436-5. DOI 10.1109/AERO50100.2021.9438269.
- SONG, Xin, SHAH, Shalin et REIF, John, 2021. Multidimensional data organization and random access in large-scale DNA storage systems. *Theoretical Computer Science*. Vol. 894, pp. 190-202. DOI 10.1016/j.tcs.2021.09.021.
- SWADE, Doron, 1998. Preserving Software in an Object-Centred Culture. In : HIGGS, Edward (éd.), *History and Electronic Artefacts*, pp. 195-206. Oxford : Clarendon Press. ISBN 0-19-823633-6.
- TOMEK, Kyle J. et al., 2019. Driving the Scalability of DNA-Based Information Storage Systems. *ACS Synthetic Biology*. Vol. 8, no 6, pp. 1241-1248. DOI 10.1021/acssynbio.9b00100.
- UBAIDUR RAHMAN, Noorul Hussain, BALAMURUGAN, Chithralekha et MARIAPPAN, Rajapandian, 2015. A Novel DNA Computing Based Encryption and Decryption Algorithm. *Procedia Computer Science*. Vol. 46, pp. 463-475. DOI 10.1016/j.procs.2015.02.045.
- UK DATA SERVICE, 2022. Recommended formats. *UK Data Service* [en ligne]. 15 août 2022. Disponible à l'adresse : <https://ukdataservice.ac.uk/learning-hub/research-data-management/format-your-data/recommended-formats/> [consulté le 15 août 2024].
- VAN DER HOEVEN, Jeffrey R., VAN DIESSEN, Raymond J. et VAN DER MEER, Kees, 2005. Development of a Universal Virtual Computer (UVC) for long-term preservation of digital objects. *Journal of Information Science*. Vol. 31, no 3, pp. 196-208. DOI 10.1177/0165551505052347.

VINCENT, John, GURLING, Hugh et MELMER, Georg, 1994. Oligonucleotides as Short as 7-Mers Can Be Used for PCR Amplification. *DNA and Cell Biology*. Vol. 13, no 1, pp. 75-82. DOI 10.1089/dna.1994.13.75.

WANG, Shaopeng et al., 2024. Data Storage Using DNA. *Advanced Materials*. Vol. 36, no 6, pp. 1-23. DOI 10.1002/adma.202307499.

WATERS, Donald et GARRETT, John, 1996. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information* [en ligne]. The Commission on Preservation and Access, 1400 16th St. Disponible à l'adresse : <https://eric.ed.gov/?id=ED395602> [consulté le 15 août 2024].

YAZDI, S. M. Hossein Tabatabaei, GABRYS, Ryan et MILENKOVIC, Olgica, 2017. Portable and Error-Free DNA-Based Data Storage. *Scientific Reports*. Vol. 7, no 1, pp. 1-6. DOI 10.1038/s41598-017-05188-1.

ZHAN, Pengfei et al., 2023. Recent Advances in DNA Origami-Engineered Nanomaterials and Applications. *Chemical Reviews*. Vol. 123, no 7, pp. 3976-4050. DOI 10.1021/acs.chemrev.3c00028.

ZHIRNOV, Victor et al., 2016. Nucleic acid memory. *Nature Materials*. Vol. 15, no 4, pp. 366-370. DOI 10.1038/nmat4594.

ZIERAU, Eld, 2017. OAIS and Distributed Digital Preservation in Practice: An exploration of Danish and other use cases that contributed to the development of the Outer OAIS–Inner OAIS Model for Distributed Digital Preservation. In : *iPRES 2017 - 14th International Conference on Digital Preservation* [en ligne]. 2017. Disponible à l'adresse : <https://phaidra.univie.ac.at/detail/o:931073> [consulté le 15 août 2024].