http://doi.org/10.7207/twgn22-02-FR

Une approche de la préservation du train de bits axée sur les risques

Paul Wheatley

Note d'orientation sur la veille technologique DPC

Mise à jour de janvier 2025

DigitalPreservationCoalition

Version 1.3, 2025

© Digital Preservation Coalition 2025 et Paul Wheatley 2025
ISSN : 2048-7916
Tous droits réservés. Aucune partie de cette publication ne peut être reproduite, stockée dans un système d'extraction ou transmise, sous quelque forme ou par quelque moyen que ce soit, sans l'autorisation écrite préalable de l'éditeur. Les droits moraux de l'auteur ont été revendiqués.
Publié pour la première fois en Grande-Bretagne en 2022 par la Digital Preservation Coalition.

© Digital Preservation Coalition 2025 et Paul Wheatley, 0000-0002-3839-3298

Ce travail est sous licence CC BY-NC-SA 4.0.

Cette note d'orientation est publiée par la Digital Preservation Coalition (DPC). La DPC est une fondation caritative internationale qui soutient la préservation numérique et aide ses membres dans le monde entier à fournir un accès à long terme et résilient aux contenus et services numériques. Outre la publication de rapports sur une série de thèmes couvrant l'état de l'art en matière de préservation numérique, la DPC soutient également ses membres par le biais d'un engagement communautaire, d'un travail de plaidoyer ciblé, d'une formation et d'un développement des ressources humaines (*workforce development*), de l'identification de bonnes pratiques et de normes, ainsi que d'une bonne gestion et d'une bonne gouvernance. Sa vision est celle d'un héritage numérique sûr.

Découvrez les publications de la DPC, y compris les dernières mises à jour et révisions, à l'adresse suivante :

https://www.dpconline.org/digipres/discover-good-practice/tech-watch-reports

Pour en savoir plus sur la DPC, le soutien qu'elle offre et la manière de devenir membre, consultez le site :

https://www.dpconline.org/about/join-us

La version 1.0 a été publiée pour la première fois en 2022.

La version 1.3, en janvier 2025, avec des références mises à jour, quelques restructurations et reformulations, des exemples supplémentaires et des ajouts de références sur les défis et les mesures d'atténuation sur le stockage dans le Cloud.

Informations sur la traduction française

La traduction française de cette note d'orientation a été réalisée dans le cadre des travaux de la Cellule nationale de veille sur les formats. Lancée en 2019, la Cellule nationale de veille sur les formats (CNVF), sous l'égide de l'association française Aristote et de son groupe de travail sur la Pérennisation de l'information numérique (PIN), regroupe à ce jour une douzaine de partenaires. Ses principaux objectifs sont la mutualisation des activités de veille sur les formats, la sensibilisation des professionnels sur le sujet, la contribution ou l'influence sur les outils associés. Elle ambitionne d'être un interlocuteur francophone reconnu dans les travaux internationaux sur ces sujets.

La note d'orientation a été traduite en français par : Thomas Ledoux (Bibliothèque nationale de France). Une première traduction automatique a été réalisée avec le logiciel DeepL puis a été revue et corrigée par le groupe de travail.

1 Introduction

Stocker le contenu numérique sans pertes est loin d'être la seule considération à prendre en compte pour assurer la préservation numérique à long terme. Mais comme l'illustre le slogan officieux de la DPC « Keep the bits », il s'agit d'une considération cruciale. La grille d'évaluation rapide de la DPC, ou DPC RAM, (DPC, 2024) indique que pour atteindre le niveau 3 « Gestion standard » pour le critère « Préservation du train de bits », « Un processus d'évaluation des risques est mis en place pour évaluer les risques liés au stockage et les mesures d'atténuation appropriées (telles que le nombre de copies, la localisation géographique, les technologies utilisées, la fréquence des contrôles d'intégrité) ». Mais qu'est-ce que cela implique en pratique ?

Cette note d'orientation explore certains des défis que doit relever une architecture de stockage conçue pour la préservation numérique à long terme. Elle examine les risques auxquels est confronté le contenu stocké à long terme et conclut par une approche simple permettant d'évaluer et de documenter les risques et les mesures d'atténuation mises en place pour y faire face.

Les professionnels de la préservation numérique peuvent trouver cette note d'orientation utile lorsqu'ils cherchent à mettre en place un stockage approprié à la conservation ou lorsqu'ils vérifient que leur stockage existant est adapté à l'objectif visé. Elle peut aider à justifier, au sein d'une organisation, les ressources nécessaires à la mise en place de mesures d'atténuation supplémentaires pour faire face aux risques identifiés en matière de stockage. En plus des « Exigences fondamentales pour un système de conservation numérique » (DPC, 2022) de portée plus large, cette note d'orientation peut être utile pour communiquer les exigences quelque peu uniques de la conservation numérique à long terme lors des contacts avec le personnel informatique.

Il convient de noter que lorsqu'on envisage une infrastructure de stockage appropriée pour la préservation, il est important de prendre en compte et de documenter de nombreuses autres exigences de stockage, telles que l'accès, l'interopérabilité et la montée en charge. Ces exigences sont considérées comme n'entrant pas dans le champ d'application de cette note d'orientation, mais elles sont décrites en détail dans le document « Preservation Storage Criteria, Version 4 » (Schaefer et al, 2024).

1.1 Qu'est-ce que la « préservation du train de bits » ?

La grille d'évaluation rapide de la DPC décrit la préservation du train de bits comme « des processus visant à garantir le stockage et l'intégrité du contenu numérique à préserver ». Les informations que nous souhaitons préserver sont codées de différentes manières, souvent à l'aide de formats de fichiers spécifiques. En fin de compte, elles sont représentées par une série de zéros et de uns. Il s'agit de chiffres binaires ou de bits. Une série de bits, représentant par exemple un fichier, est souvent appelée « train de bits » ou « flux binaire ». La préservation du train de bits se concentre sur la permanence de cette série de bits. Elle ne traite pas de la manière dont les informations sont encodées dans ces bits, ni, ce qui est plus crucial pour la préservation, de la manière dont une série de bits peut être décodé en informations utiles (connue sous le nom distinct mais connexe de « Préservation des contenus » dans le cadre de la grille d'évaluation rapide de la DPC).

1.2 Comprendre les exigences de stockage de la préservation numérique

Bien que les risques auxquels sont confrontés les besoins de stockage du monde de la préservation numérique ne soient pas totalement différents de ceux d'un environnement informatique plus classique, les exigences différentes des spécialistes de la préservation peuvent nécessiter l'adoption d'une approche différente. Rosenthal et al (2005) notent que « bon nombre de ces menaces ne sont pas propres aux systèmes de préservation numérique, mais leur mission spécifique et leurs horizons

à très long terme incitent ces systèmes à envisager les menaces différemment des systèmes plus conventionnels ».

Un mode de stockage informatique typique peut être conçu pour fournir des services résilients (i.e. avec une faible indisponibilité) aux utilisateurs dès maintenant, avec une certaine facilité de sauvegarde et de récupération à court terme. La préservation numérique à long terme se préoccupe généralement moins des interruptions mineures du fonctionnement immédiat ou de l'accès au contenu (temps d'arrêt), mais doit pouvoir garantir qu'aucun (ou peu) de contenu ne sera perdu sur le véritable long terme (souvent défini comme 100 ans ou plus) (Prater, 2018).

Une réponse purement informatique probable à une attaque importante par rançongiciel (*ransomware*) pourrait consister à mettre en place de nouveaux services métiers à partir de zéro, probablement en tirant parti des technologies basées sur le cloud. Mais cette approche de reprise après sinistre axée sur les services métiers n'est pas aussi pertinente dans le monde de la préservation numérique, où la conservation des données est primordiale. Pour parler franchement, si tout le contenu numérique qui a été patiemment collecté pendant des décennies est perdu, il n'y a pas de reprise après sinistre. Il n'y a qu'un sinistre.

2 Une approche basée sur les risques

Le texte suivant est un simple guide pour évaluer les risques de stockage rencontrés dans un cas particulier, et pour envisager un ensemble approprié de mesures d'atténuation de ces risques.

La réponse à la question triviale « Combien de copies dois-je conserver ? » est classiquement mais sans doute simplement « 3 » (NDSA 2019). Il faut plus de détails pour comprendre l'efficacité de la réduction des risques offerte par la redondance multi-copies en question. Par exemple, conserver 3 copies (lorsque la plupart des utilisateurs ne peuvent pas accéder aux 3 copies) réduira le risque de perte de données due à une erreur humaine. Mais si ces 3 copies se trouvent dans le même bâtiment, elles n'offrent qu'une faible protection contre les risques d'incendie.

Il n'y a pas de solution unique qui convienne à tous les contextes. Une organisation peut avoir un profil de risque différent d'une autre. L'appétit pour le risque, la valeur du contenu et les ressources disponibles pour l'atténuation peuvent également varier selon chaque organisation.

La présente note d'orientation est donc conçue pour guider le professionnel dans le processus de réflexion et d'évaluation des risques liés au stockage de leur propre contenu numérique, d'une manière adaptée à *leur* organisation et à *leurs* exigences.

2.1 Quand les pertes se produisent-elles ?

Les organisations sont souvent réticentes à partager les détails de perte de données, mais un thème commun dans ceux qui ont été publiés semble être la conjonction de multiples problèmes se produisant en même temps. Il peut s'agir d'une erreur humaine, d'une panne de courant, d'une catastrophe naturelle ou d'origine humaine, ou d'un comportement inattendu du logiciel dû à des bugs (The Register, 2017a ; The Register, 2017b). Les données sont encore plus en danger quand elles sont en déplacement, au cours des processus de gestion, de transfert de données ou de migrations pour le rafraîchissement des supports. Parmi les autres facteurs courants, on peut citer le fait de ne pas mener à bien ou de ne pas vérifier complètement des processus tels que le contrôle d'intégrité, l'application de correctifs aux logiciels ou la sauvegarde du contenu. Une leçon importante à tirer de cette situation est que l'établissement de processus pour atténuer les risques liés au stockage est, à lui seul, insuffisant pour assurer la préservation. Les processus d'atténuation

doivent être eux-mêmes soigneusement contrôlés, validés et, idéalement, évalués de manière indépendante pour garantir leur efficacité continue.

Les rançongiciels continuent de représenter un risque majeur pour la survie des contenus numériques. L'attaque très médiatisée contre la *British Library* (Wikipedia, 2023) et l'absence de consensus sur la manière de lutter contre la vague croissante d'attaques par rançongiciel (The Register, 2024) soulignent la nécessité permanente de préserver les données numériques afin d'atténuer cette menace par des moyens qui vont au-delà des mesures de cybersécurité habituelles.

L'avènement de l'externalisation du stockage a réduit le risque de perte due à une défaillance matérielle, mais a entraîné de nouvelles menaces pour les données. En 2024, *UniSuper*, un gestionnaire de fonds australien, a découvert que « l'ensemble de son abonnement d'infrastructure avait été supprimé » (Mellor 2024). Les deux copies de ses données dans Google Cloud ont été perdues, mais la plupart des données ont finalement été récupérées à partir d'une troisième copie stockée chez un autre fournisseur cloud.

L'enquête NDSA 2021 sur la fixité (NDSA, 2021) a interrogé des organisations opérant dans le domaine de la préservation à long terme. Elle semble indiquer que peu d'organisations ont subi des échecs fréquents de contrôle d'intégrité et que peu d'entre eux étaient liés à un stockage dédié à la préservation numérique. Aucun chiffre n'est fourni sur l'ampleur ou l'impact de ces défaillances, mais les détails sur la nature de la cause et de la correction impliquent que nombre d'entre elles sont de faible ampleur. Cela permet de croire que les mesures d'atténuation des risques fonctionnent avec un certain degré d'efficacité. La poursuite de la collecte de données plus riches dans ce domaine devrait s'avérer inestimable pour éclairer les approches de préservation appropriées et justifier les investissements nécessaires à leur mise en œuvre, avec, espérons-le, un impact économique et écologique minimal (Stokes, 2022).

2.2 Pourquoi utiliser une approche du stockage fondée sur les risques ?

La préservation du train de bits est un élément fondamental pour garantir que l'information numérique puisse être préservée sur le long terme et, en fin de compte, qu'on puisse y accéder en réalisant sa valeur. Il est donc essentiel de veiller à ce que les risques soient identifiés, compris et gérés de manière appropriée. Cette approche présente toutefois d'autres avantages.

Documenter les menaces et les mesures d'atténuation est largement reconnu comme étant une bonne pratique. Une évaluation formelle des risques aboutira à une documentation du processus et du résultat, fournissant la preuve des activités de planification de la préservation qui pourraient être nécessaires pour la certification des archives/de la préservation. La norme de certification Core Trust Seal pose la question suivante : « Les techniques de gestion des risques sont-elles utilisées pour informer la stratégie ? » et exige des pièces justificatives documentées pour être conforme (Core Trust Seal, 2022).

Une évaluation formelle des risques peut également servir de preuves utiles lorsqu'il s'agit d'obtenir des ressources pour mettre en œuvre des mesures complémentaires d'atténuation des risques. La communication des raisons qui sous-tendent les exigences quelque peu particulières de la préservation numérique à long terme reste un défi organisationnel de taille. Mettre en évidence les lacunes en termes de capacité de préservation et l'impact d'une éventuelle perte de contenu, des coûts financiers et d'atteinte à la réputation peut être un moyen efficace de rallier les cadres supérieurs à votre cause.

2.3 Étapes de l'application d'une évaluation des risques pour le stockage de la préservation numérique

Un simple processus d'évaluation des risques sera suffisant pour guider l'examen des risques de préservation acceptables (ou inacceptables), mais il doit couvrir l'ensemble du champ d'application et constituer une évaluation honnête des risques, de leur probabilité et de leurs impacts. La norme ISO 27001 sur la sécurité de l'information (Wikipedia, 2022) peut fournir des conseils utiles pour définir et documenter une approche d'évaluation des risques. De nombreuses organisations ont leur propre processus de gestion des risques qui peut être utilement mis en œuvre pour ce faire. Sinon, les étapes clés d'un processus simple d'évaluation des risques sont décrites ci-dessous :

1. Identifier et consigner le champ d'application de l'évaluation des risques, en détaillant en particulier le contenu numérique auquel elle s'appliquera.
2. Identifier les risques significatifs en rapport avec le champ d'application défini.
3. Attribuer une note à la probabilité d'occurrence de chaque risque et à l'impact qu'il aura s'il se produit. Ces scores peuvent être multipliés pour générer un score initial pour chaque risque.
4. Documenter les mesures d'atténuation des risques en place dans votre organisation et fournir un score ajusté pour chaque risque qui prend en compte les mesures d'atténuation.
5. Tenir compte de l'appétence de votre organisation pour les scores de risque ajustés qui ont été générés. Il peut être utile de consulter une série de parties prenantes internes, la direction générale et éventuellement des conseillers externes tels que la DPC ou des organisations homologues.
6. Documenter toute mesure d'atténuation supplémentaire jugée nécessaire pour traiter les risques en suspens.

Il n'existe pas de réponse unique et correcte à la question de savoir quelles mesures d'atténuation des risques sont appropriées pour une organisation donnée. Toute mesure d'atténuation particulière peut réduire le risque de préservation, mais elle entraînera probablement aussi un coût financier et éventuellement un coût environnemental. Il peut être nécessaire de prendre en compte le caractère unique et la valeur du contenu à préserver (ou, à l'inverse, le coût financier ou de réputation de la perte du contenu) et le niveau de perte qui pourrait être acceptable (Pendergrass et al, 2019), afin d'identifier un niveau de risque acceptable. Par conséquent, il peut être utile de développer des profils de risque pour différentes collections ou niveaux d'engagement de préservation des documents, comme dans cet exemple des bibliothèques de l'université de Penn State (2021). Les systèmes de préservation numérique offrent de plus en plus de possibilités aux utilisateurs d'adapter les profils de stockage à des fonds particuliers.

Le tableau suivant fournit un résumé des risques de stockage courants et des mesures d'atténuation typiques qui peuvent y être associées, mais d'autres risques peuvent être pertinents dans votre situation :

Risques/menaces liés au stockage	**Mesures d'atténuation potentielles**
Dégradation des bits / perte ou endommagement du contenu	• Répliquer le contenu pour créer des copies redondantes • Mettre en œuvre le contrôle d'intégrité et la réparation

Défaillance du matériel de stockage	• Surveiller, gérer et réparer/remplacer le matériel de stockage • Mettre en œuvre le contrôle d'intégrité et la réparation
Obsolescence des supports de stockage et du matériel	• Planifier et mettre en œuvre le rafraîchissement/remplacement des supports de stockage/du matériel avant leur fin de vie
Suppression accidentelle / erreur humaine / dommage malveillant par le personnel	• Répliquer le contenu pour créer des copies redondantes • Assurer un contrôle rigoureux de l'accès en écriture et le principe du moindre privilège • Mettre en œuvre le contrôle d'intégrité et la réparation • Établir un processus de gestion des modifications et des suppressions légitimes de contenu • Documenter/auditer toutes les actions entraînant une altération du contenu
Dommages malveillants par des tiers	• Mettre en œuvre des mesures de cybersécurité • Reproduire le contenu pour créer des copies redondantes • Conserver des copies du contenu avec de modes de gestion différenciés • Créer une copie hors ligne du contenu • Mettre en œuvre le contrôle d'intégrité et la réparation
Défaillance par mise en commun (un seul point de défaillance matérielle ou logicielle affectant toutes les copies répliquées)	• Utiliser une combinaison de technologies matérielles et logicielles
Catastrophe naturelle / d'origine humaine	• Répliquer le contenu sur des sites géographiquement séparés présentant des profils de risque différents. • Établir une politique et une procédure de reprise après sinistre
Défaillance ou fermeture d'un fournisseur de stockage tiers	• Établir un plan d'action en cas de fermeture inattendue • Éviter la dépendance à l'égard d'un seul fournisseur tiers (p. ex. fournisseur de services en ligne) • Garantir un accès indépendant au stockage en cloud revendu par le fournisseur tiers • Utiliser des dispositifs de dépôt en séquestre
Ne pas mettre en œuvre les processus d'atténuation des risques (ci-dessus) ou ne pas vérifier qu'ils fonctionnent efficacement	• Documenter les procédures de gestion du stockage • Tester et valider les mesures d'atténuation • Fournir des rapports clairs sur la mise en œuvre des processus d'atténuation des risques à l'organe de gouvernance active • Mettre en place un audit/une certification indépendant(e) des processus et procédures de préservation à long terme

Un thème récurrent de ces mesures d'atténuation est la nécessité d'assurer la *diversité* au sein d'une infrastructure de stockage. La diversité vise à éliminer ou, à tout le moins, à réduire au minimum les points de défaillance uniques (SPOF), notamment les personnes, les logiciels, le matériel, l'emplacement géographique ou le fournisseur de services.

De manière anecdotique, la communauté de la préservation numérique a souvent identifié l'erreur humaine comme le risque de préservation numérique le plus important. Réfléchissez à la manière dont l'erreur humaine pourrait jouer un rôle dans la probabilité et l'impact de tous les risques décrits ci-dessus. La menace croissante des attaques par rançongiciel (*ransomware*) est également susceptible d'être considérée comme l'un des risques les plus critiques auxquels il faut faire face.

Les ressources suivantes fournissent des informations complémentaires précieuses sur la gestion des risques dans le contexte du stockage de préservation :

- Le *Usage Guide for the Preservation Storage Criteria* décrit une classification utile des types de risques et propose une analyse approfondie de la gestion de risques, de l'intégrité des bits et de l'indépendance des copies de stockage (Schaefer et al., 2019).
- Une exploration approfondie de la gestion de risques pour la préservation numérique est fournie par Pennock (Pennock, 2024).
- L'outil *Digital Archiving Graphical Risk Assessment Model*, ou DiAGRAM (National Archives, 2020) utilise une méthode statistique appelée réseau bayésien pour produire un modèle graphique des risques de préservation numérique, qui met l'accent sur le stockage de la préservation numérique.

3 Appréhender un environnement changeant sur l'atténuation des risques

Diverses menaces pèsent sur nos contenus numériques, quelque peu dissimulées dans les applications, les *middlewares* et les services tiers dont nous dépendons pour gérer nos contenus numériques. L'externalisation de nos fonctions de préservation peut présenter de grands avantages, mais elle modifie également le profil des menaces qui pèsent sur nos contenus numériques. Cette section examine certains des aspects que nous connaissons – et ceux que nous ignorons – au sujet de ces technologies et de ces services en constante évolution, ainsi que les approches que cette communauté commence à mettre en œuvre pour atténuer ces menaces nouvelles et émergentes.

3.1 Une image floue de la diversité du stockage, de la réplication et de la fourniture de services

Les services de stockage en ligne (*cloud storage*) offrent une multitude d'avantages potentiels pour le stockage de contenu à long terme. Cependant, un certain nombre de risques potentiels et de préoccupations ont été soulevés à propos du stockage en ligne, malgré son adoption rapide au sein de la communauté de la préservation numérique. Quel degré de confiance peut-on accorder à l'externalisation non seulement du stockage, mais aussi du contrôle d'intégrité du stockage et d'autres processus tels que le rafraîchissement des supports ? Rosenthal (2019) note en 2019 que « le contrôle de l'intégrité des données stockées dans un service en ligne sans faire confiance au service dans une certaine mesure est un problème difficile pour lequel aucune solution entièrement satisfaisante n'a été publiée. »

L'omniprésence et l'échelle du stockage en ligne au-delà du domaine de la préservation numérique suggèrent une technologie résistante, qui a probablement été testée de manière beaucoup plus rigoureuse que de nombreux autres points de risque dans une architecture de préservation numérique. La communauté de la préservation numérique, naturellement peu encline à prendre des risques, a toutefois adopté jusqu'à présent diverses approches :

- La *Wellcome Library* a fait le choix d'une approche entièrement basée sur le Cloud, mais a introduit de la diversité dans son stockage en utilisant deux fournisseurs de services en ligne différents (Chan, 2021). Chan note que « le niveau d'intégrité et de sécurité des données qu'ils (le cloud) fournissent va bien au-delà de tout ce que nous pourrions construire en interne. Nous faisons confiance à leur processus de vérification, et nous ne réalisons pas de contrôle supplémentaire ». Néanmoins, la *Wellcome* réalise un contrôle de relecture des données transférées dans le cloud, de manière à valider le succès du versement.
- Les Archives Nationales (UK) ont externalisé leur stockage et leur plateforme de préservation à une partie tierce, avec deux copies stockées dans le cloud.
- Les Archives Nationales (Royaume-Uni) ont externalisé leur stockage et leur plateforme de conservation à un tiers, avec deux copies conservées dans le cloud. Il est essentiel de noter qu'elles conservent également une troisième copie « de réserve » sur place (Daly, 2024), sous une forme basée sur le format *Oxford Common File Layout* (Jefferies et Woods, 2024), qui est compréhensible de manière indépendante. Cela ajoute de la diversité afin d'atténuer les menaces liées au stockage, et le découplage d'avec le service géré par un tiers offre une certaine agilité face à l'évolution rapide des technologies et des fournisseurs de services.
- La Bibliothèque nationale d'Écosse utilise une combinaison de stockage sur site et en ligne, tout en appliquant un contrôle d'intégrité complet ou par échantillonnage des différents points de stockage (Hibberd, 2020).
- Le *Natural Environment Research Council* a totalement évité les services en ligne, ce qui lui permet de contrôler totalement ses fonctions de contrôle d'intégrité (NDSA, 2021, p.68).

3.2 Comment vérifier l'efficacité de la préservation quand le stockage et le contrôle d'intégrité sont externalisés ?

La plupart des fournisseurs de stockage en ligne incluent des contrôles d'intégrité dans leurs services de stockage, mais ils ne divulguent généralement pas la manière dont ces contrôles sont effectués. Les affirmations relatives à une durabilité élevée restent difficiles à vérifier. Environ la moitié des personnes interrogées dans le cadre de l'enquête *NDSA 2021 Fixity Survey* (NDSA, 2021, p. 44) qui utilisaient le stockage en ligne ont déclaré avoir reçu certaines informations relatives à l'intégrité de la part de leurs fournisseurs. Environ un tiers des répondants qui ont reçu des informations sur l'intégrité n'ont pas pu les utiliser, pour diverses raisons. Certains éléments indiquent que des tiers sont à l'écoute des commentaires des utilisateurs et améliorent la prise en charge de la vérification de l'intégrité, au moins pendant le transfert vers et depuis le cloud (Stormacq, 2024).

La vérification indépendante de l'intégrité des données stockées en ligne par recalcul des sommes de contrôle a été démontrée avec succès par l'université de l'Illinois à Urbana-Champaign, malgré des défis pratiques et économiques (Rimkus et Schmitt, 2024). Dans un monde idéal, les organisations n'auraient pas besoin de payer pour dupliquer la vérification d'intégrité revendiquée par les fournisseurs de services en ligne et ne paieraient pas deux fois pour une fonction qu'elles ont externalisée. Ayant acquis une certaine confiance dans ses fournisseurs grâce à des contrôles d'intégrité complets et indépendants au cours des années précédentes, l'université de l'Illinois a depuis mis en place une politique de vérification des nouveaux contenus et d'échantillonnage aléatoire des contenus stockés, au lieu de procéder à de nouveaux contrôles d'intégrité complets. « L'objectif ici est de croire que ce niveau de durabilité des fichiers se maintiendra, mais de vérifier cette durabilité à un rythme modéré afin de détecter les failles dans le stockage et la gestion des fichiers, si celles-ci devaient apparaître. »

La communauté a commencé à discuter, et dans certains cas à mettre en œuvre, des approches alternatives et/ou complémentaires aux contrôles d'intégrité complets. Cela a été en partie motivé par l'identification de menaces associées en particulier au stockage en ligne externalisé, comme l'exemple de suppression de compte mentionné au point 2.1 ci-dessus. Il s'agit notamment des menaces suivantes :

- Surveiller les changements dans les inventaires des éléments stockés en ligne ou effectuer des vérifications complètes des manifestes de fichiers qui ne vont pas jusqu'à des vérifications d'intégrité nécessitant une puissance de calcul importante (et donc coûteuses en termes de *cloud computing*). C'est un bon moyen de détecter les problèmes dus à des erreurs humaines ou à des problèmes logiciels, tels que des erreurs dans les flux de travail automatisés, tout en faisant confiance à la plateforme de stockage pour conserver les données stockées.
- Vérifier si un service tiers est toujours opérationnel en demandant quotidiennement un petit nombre de fichiers (Altman et Landau, 2024). Identifier rapidement qu'un service de stockage ne fonctionne plus permet de prendre des mesures correctives ou de mettre en place des alternatives.
- Séparer la gestion des contrats de plusieurs instances cloud détenues auprès de différents fournisseurs entre différents services organisationnels et responsables afin de réduire le risque d'erreurs comptables ou de facturation entraînant la fermeture accidentelle de toutes les copies en même temps.
- Vérifier automatiquement la facturation des fournisseurs en ligne via une API. Un changement soudain et important dans la facturation régulière des services en ligne peut indiquer un changement catastrophique dans ce qui est (ou n'est pas) stocké.

En fin de compte, l'ampleur de l'utilisation des services de stockage en ligne et l'absence de problèmes connus liés à la détérioration des bits suggèrent qu'il n'est pas nécessaire de reproduire les contrôles d'intégrité complets effectués par les fournisseurs de services cloud. Cette communauté n'a toutefois pas encore convenu d'un niveau de confiance et d'atténuation des risques pour l'utilisation des services en ligne à des fins de préservation à long terme.

3.3 Points de défaillance uniques (SPOF) cachés ?

Hockx-Yu et Brewer (2021) s'inquiètent des risques potentiels présents dans les systèmes d'intermédiation du stockage tels que les passerelles des stockages en ligne comme *AWS Storage Gateway*, et les passerelles de stockage sur bande comme le *BlackPearl Converged Storage System* de Spectra Logic. Les systèmes d'intermédiation du stockage peuvent fournir un accès pratique à des emplacements de stockage multiples et apparemment divers, mais ils peuvent également introduire des points de défaillance uniques et des points d'attaque externes uniques. « Les systèmes d'intermédiation du stockage remettent directement en question la notion de redondance… » Ils recommandent « … d'accroître la sensibilisation et d'approfondir la compréhension de ces systèmes d'intermédiation, en particulier la façon dont ils peuvent devenir le point unique de défaillance conduisant à la perte de données ou à la préservation numérique ».

Les applications commerciales ou libres des systèmes de préservation numérique sont de plus en plus utilisées pour gérer et fournir le stockage, le contrôle d'intégrité et une variété d'autres services pertinents pour cette note d'orientation. Il convient de tenir compte du fait que ces systèmes peuvent présenter des points de défaillance uniques, indépendamment de la réplication et des autres mesures d'atténuation employées au niveau du stockage. Des exemples de pertes dues à des

bugs logiciels dans les systèmes de préservation ont été vécus. Les professionnels chargés de la préservation doivent continuer à interpeller les fournisseurs de systèmes de préservation dans ce domaine et collaborer avec eux pour signaler et traiter tous les problèmes potentiels qui pourraient être identifiés.

4 Conclusion

La conception et la mise en œuvre d'infrastructure de stockage pour la préservation numérique à long terme sont souvent influencées ou dirigées par une série de facteurs sans rapport avec les préoccupations liées à la conservation des données sur de longues périodes. Des questions telles que les ressources limitées, l'aspect pratique, les politiques organisationnelles d'externalisation des technologies de l'information et bien d'autres encore peuvent détourner l'attention d'une question essentielle : une infrastructure de stockage particulière est-elle suffisante pour préserver les données à long terme ? Un simple processus d'évaluation des risques peut être une approche utile pour documenter les informations clés nécessaires pour répondre à cette question et identifier où une atténuation supplémentaire des risques peut être nécessaire. Cette approche reste toujours valable alors même que l'externalisation des fonctions de préservation progresse rapidement. L'évaluation des risques continue de montrer que la diversité des modes de stockage est hautement souhaitable. C'est cette diversité qui est essentielle pour atténuer la grande variété de menaces auxquelles notre contenu sera confronté à long terme.

5 Bibliographie

Addis, M. (2020) *Which checksum algorithm should I use?* Disponible à l'adresse suivante : http://doi.org/10.7207/twgn20-12

Altman, M and Landau, R. (2024) *Selecting Efficient and Reliable Preservation Strategies: Modeling Long-term Information Integrity Using Large-scale Hierarchical Discrete Event Simulation*. IJDC. Disponible à l'adresse suivante : https://doi.org/10.2218/ijdc.v18i1.743

Chan, A. (2021) *Our approach to digital verification*. Disponible à l'adresse suivante : https://web.archive.org/web/20220809134815/https://stacks.wellcomecollection.org/our-approach-to-digital-verification-79da59da4ab7?gi=f955d8d0d5c1

Core Trust Seal (2022) *Core Trust Seal*. Disponible à l'adresse suivante : https://web.archive.org/web/20220802114709/https://www.coretrustseal.org/

Daly, S (2024) *A decoupled Custodial Copy for cloud-based Digital Preservation Systems*. Disponible à l'adresse suivante : https://doi.org/10.5281/zenodo.13647419

DPC (2021) *DPC Rapid Assessment Model Version 3*. Disponible à l'adresse suivante : https://web.archive.org/web/20250131231437/https://www.dpconline.org/digipres/implementdigipres/dpc-ram

DPC (2022) *Core requirements for a digital preservation system*. Disponible à l'adresse suivante : https://web.archive.org/web/20220810111732/https://www.dpconline.org/digipres/implement-digipres/core-requirements-for-a-digital-preservation-system

Jefferies, N and Woods, A. (2024) *The Oxford Common File Layout*, Github, Disponible à l'adresse suivante : https://github.com/OCFL

Mellor, C. (2024) *Google Cloud deleted a large customer's infrastructure. Blocks and Files*. Disponible à l'adresse suivante : https://web.archive.org/web/20250131234033/https://blocksandfiles.com/2024/05/14/googlecloud-unisuper/

National Archives (2020) *DiAGRAM - The Digital Archiving Graphical Risk Assessment Model*. Disponible à l'adresse suivante : https://web.archive.org/web/20220809153306/https://nationalarchives.shinyapps.io/DiAGRAM/

NDSA (2019) *2019 Storage Infrastructure Survey*. Disponible à l'adresse suivante : https://doi.org/10.17605/OSF.IO/UWSG7

NDSA (2021) *2021 Fixity Survey*. Disponible à l'adresse suivante : https://doi.org/10.17605/OSF.IO/2QKEA

Pendergrass, K. L. Sampson, W. Walsh, T. and Alagna, L. (2019) *Toward Environmentally Sustainable Digital Preservation*, American Archivist, Volume 82, Issue 1. Disponible à l'adresse suivante : https://web.archive.org/web/20220804114356/https://meridian.allenpress.com/american-archivist/article/82/1/165/432804/Toward-Environmentally-Sustainable-Digital

Penn State University Libraries (2021) *Policy UL-AD19 Digital Preservation Policy*. Disponible à l'adresse suivante : https://web.archive.org/web/20220804123736/https://libraries.psu.edu/policies/ulad-19

Pennock, M. (2024) *Disentangling Digital Preservation Risk: An Interdisciplinary Exploration and Solution*. Disponible à l'adresse suivante : https://dx.doi.org/10.15132/20000457

Prater S. (2018) *How to Talk to IT about Digital Preservation*, Journal of Archival Organization, Disponible à l'adresse suivante : https://web.archive.org/web/20210520101609/https://minds.wisconsin.edu/bitstream/handle/1793/78844/How%20to%20Talk%20to%20IT%20about%20Digital%20Preservation.pdf?sequence=3&isAllowed=y

The Register (2017) *GitLab.com melts down after wrong directory deleted, backups fail*. Disponible à l'adresse suivante : https://web.archive.org/web/20220729152515/https://www.theregister.com/2017/02/01/gitlab_data_loss/

The Register (2017) *KCL external review blames whole IT team for mega-outage, leaves managers unshamed*. Disponible à l'adresse suivante : https://web.archive.org/web/20220729152543/https://www.theregister.com/2017/02/23/kcl_external_review/

The Register (2024) *What do ransomware and Jesus have in common? A birth month and an unwillingness to die*. Disponible à l'adresse suivante : https://www.theregister.com/2024/12/24/ransomware_in_2024/

Rimkus, K.R. and Schmitt, G. (2024). *File Fixity in the Cloud: Policy, Business, and Technical Considerations*. iPRES 2024. Disponible à l'adresse suivante : https://doi.org/10.21428/5676bf2d.3d7946ac

Rosenthal et al. (2005) *Requirements for Digital Preservation Systems*, DLib November 2005, Volume 11, Number 11. Disponible à l'adresse suivante : https://web.archive.org/web/20220423182212/http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

Rosenthal D. (2019) *DSHR's Blog: Cloud for Preservation*. Disponible à l'adresse suivante : https://web.archive.org/web/20220804151256/https://blog.dshr.org/2019/02/cloud-for-preservation.html

Schaefer et al. (2019) *Usage Guide for the Preservation Storage Criteria*. Disponible à l'adresse suivante : https://osf.io/4cvqa

Schaefer et al. (2015) *Digital Preservation Storage Criteria*. Disponible à l'adresse suivante : https://doi.org/10.17605/OSF.IO/SJC6U

Stokes, P. (2022). *Catastrophic data loss is going to cost us how much….?!.* Disponible à l'adresse suivante : https://web.archive.org/web/20220830114430/https://www.dpconline.org/blog/stokes-cost-of-catastophic-data-loss 7

Stormacq, S. (2024). *Introducing default data integrity protections for new objects in Amazon S3*.

Amazon. Disponible à l'adresse suivante : https://aws.amazon.com/blogs/aws/introducing-default-data-integrityprotections-for-new-objects-in-amazon-s3/

Wikipedia (2022) *ISO/IEC 27001*, Disponible à l'adresse suivante : https://web.archive.org/web/20220804122211/https://en.wikipedia.org/wiki/ISO/IEC_27001

Wikipedia (2023) *British Library cyberattack*, Disponible à l'adresse suivante : https://en.wikipedia.org/wiki/British_Library_cyberattack