

# LA DEMARCHE QUALITE AU CINES POUR LA PRESERVATION A LONG-TERME DES DONNEES NUMERIQUES

*Olivier Rouchon*

CINES – Centre Informatique National de l'Enseignement Supérieur

olivier.rouchon@cines.fr

## 1. INTRODUCTION

L'information créée, manipulée, traitée, stockée, échangée, diffusée dans notre société est une information majoritairement sous forme numérique. Cette forme de représentation de l'information est incroyablement puissante et le coût de son stockage est de plus en plus faible ; il est à présent possible de la préserver sans la moindre altération, et des outils existent pour créer des documents complexes et pour y rechercher les informations utiles. Et pourtant, en arrière plan de ces avantages immenses, se cache un risque majeur, celui d'une grande vulnérabilité au temps, explicite et avérée.

Cette vulnérabilité tient à la réalisation inéluctable, si aucune action préventive n'est prise, d'un ou de plusieurs risques liés à la nature même de l'information, à savoir :

- La connaissance perdue du contenu des objets numériques : pour pouvoir gérer et retrouver une information, il est nécessaire de disposer des détails permettant de la signaler et de la localiser. Sans cette description, la possibilité de retrouver l'information repose sur la mémoire de celui qui l'a créée.

- L'impossibilité de lire le format des fichiers contenant l'information : une information sous forme numérique n'est rien d'autre qu'une suite de 0 et de 1, et il faut se référer à un format standard du marché pour pouvoir en restituer le contenu informationnel. Or, ces standards existent en très grand nombre et sont aussi en constante évolution ; certains formats largement utilisés il y a quelques années ont aujourd'hui disparu.

- La dégradation et le vieillissement des supports de stockage : un dispositif de lecture complexe impliquant du matériel et du logiciel est indispensable pour l'accès à une donnée numérique. Ce dispositif peut s'altérer dans le temps, à cause d'effets physiques ou chimiques, tels que les rayonnements électromagnétiques, la chaleur ou encore la poussière, rendant impossible la lecture correcte de l'information.

- La disparition des matériels ou logiciels de lecture : la pérennité d'un tel dispositif de lecture dépend de celle de son fabricant ou de son bon vouloir, le bon vouloir de l'éditeur de logiciel pouvant être de vous garder sous sa dépendance. L'indisponibilité d'un seul élément au sein du

dispositif de lecture peut compromettre définitivement un patrimoine de données. Chacun sait à quelle vitesse évoluent les technologies numériques et à quel rythme les nouvelles éliminent les anciennes. Nous sommes confrontés ici à la faible pérennité des supports de données, à la disparition pure et simple des technologies de stockage les unes après les autres. Une telle disparition signifie en pratique que les matériels utilisés par cette technologie ne sont plus fabriqués et disparaissent du marché avec pour conséquence un accroissement exponentiel des coûts de maintenance des équipements existants.

L'échelle de temps est donc ici un paramètre majeur eu égard au problème posé. Si on se situe à un horizon de l'ordre de dix ans, le problème est - relativement - simple à traiter. En effet, un stockage informatique de bonne qualité et sécurisé garantit contre la perte accidentelle du document. L'évolution des technologies n'aura sans doute pas conduit à un bouleversement tel que le document sera devenu irrémédiablement illisible. Et enfin, la communauté des utilisateurs potentiels du document sera vraisemblablement assez proche, scientifiquement et culturellement, de celle qui aura créé le document dix ans plus tôt.

Si on se place à un horizon de l'ordre de trente ans ou plus, rien de tout cela n'est plus assuré si personne ne s'est préoccupé d'accompagner le document dans le temps. C'est bien cet horizon du très long terme qui constitue le cœur du défi de l'archivage numérique pérenne, à savoir atténuer l'impact des quatre risques évoqués précédemment le jour où ils se réalisent, et ce sont bien les effets d'une démarche d'assurance qualité qui permettront de garantir l'intelligibilité et l'accessibilité des documents numériques. Les méthodes essentielles à mettre en œuvre sont maintenant éprouvées :

- L'utilisation de métadonnées et d'identifiants persistants pour la capacité à garder la connaissance des contenus,

- Le choix de formats de fichier durables pour en garder la maîtrise et la capacité à les migrer vers de nouveaux formats (lorsque la conversion est la stratégie de préservation),

- La gestion proactive du vieillissement des médias de stockage pour la capacité à conserver correctement le train de bits constituant les fichiers et les migrer vers de nouveaux supports,

- La veille technologique permanente et l'anticipation des changements technologiques.

L'approche qualité au travers de bonnes pratiques dans ce domaine peut être abordée sous deux angles, technique et organisationnel. C'est cette initiative qui a été mise en place au CINES, pour la préservation du patrimoine scientifique de la communauté Enseignement Supérieur et Recherche.

## 2. LA QUALITE TECHNIQUE

L'approche technique de la qualité pour la préservation de documents numériques recouvre l'ensemble des procédures visant à garantir un haut niveau de qualité de l'objet numérique lui-même. Elle peut se décliner en trois niveaux.

### 2.1. La qualité des métadonnées

Les métadonnées sont les données qui permettent de préserver les informations décrivant les objets numériques ; on parle alors d'informations de pérennisation (description du contenu de l'information, de sa provenance ou sa source, et son historique) et d'informations de représentation (aspect technique de l'information, de sa structure ou sa forme, et de ses droits d'accès). Plusieurs types de contrôle peuvent être effectués, lesquels pourront être facilités par l'adoption de standards de métadonnées : en effet, les métadonnées normalisées sont décrites dans un référentiel qu'il est facile d'utiliser comme base de vérification pour s'assurer du niveau de qualité. L'éventail de tels jeux de métadonnées est très large, du plus générique pour la description de ressources numériques (ex. Dublin Core, ISO 15836, utilisé au CINES pour détailler l'information associée à un objet numérique à préserver) aux plus spécifiques à un domaine précis comme le commerce électronique (ex. ebXML, ISO 15000) ou les données géographiques (ISO 19115), en passant par les métadonnées techniques de préservation (PREMIS, METS) ou encore les métadonnées administratives pour la propriété intellectuelle et les droits d'auteur (MPEG-21).

### 2.2. La qualité des formats de fichiers

Pour pouvoir être lisibles et convertibles dans le temps, les fichiers doivent respecter scrupuleusement les spécifications de leur format ; des outils libres (Jhove, DROID), ainsi que des bases de données spécifiques aux formats (PRONOM) faisant autorité dans le domaine de l'archivage électronique permettent l'identification, la validation et la caractérisation des fichiers et assurent leur éligibilité à une conservation à long terme. Afin de permettre le contrôle de la qualité d'un fichier, celui-ci doit donc être dans un format identifié et vérifiable. Il s'agit par conséquent de privilégier des formats de fichiers maîtrisés, c'est-à-dire dont les spécifications sont publiées (ex. PNG, ISO 15948) et normalisées si possible (ex. PDF, ISO 32000-

1), et surtout largement utilisés par la communauté des producteurs d'information. Dans cette optique, la plateforme de préservation numérique du CINES prend en charge une liste restreinte de formats de fichiers (une quinzaine au total à l'heure actuelle), et intègre ces logiciels libres pour effectuer toute une série de contrôle qualité sur les fichiers déposés avant leur archivage à long terme.

### 2.3. La qualité du stockage

La bonne conservation du train de bits qui compose les fichiers contenant les données à préserver est cruciale ; elle constitue le niveau le plus bas, le plus proche du média physique utilisé pour entreposer l'information. La première stratégie consiste à effectuer des copies multiples d'un même document, si possible dans un nombre supérieur à deux. La détermination du nombre de copies peut être problématique, notamment sur un plan financier, car lorsque la volumétrie des documents à préserver augmente, les coûts de stockage augmentent en proportion, réduisant potentiellement le nombre de copies qui peut être réalisé. Il convient également de d'utiliser plusieurs technologies de support de stockage, idéalement différentes, pour éviter d'hypothétiques erreurs de conception ou structurelles qui pourraient apparaître au fil du temps. Enfin, un audit régulier de toutes les répliques permettra de détecter les dommages éventuels, et d'effectuer les actions correctives si besoin est.

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

La qualité technique est donc incontournable, mais n'a de sens que si l'organisation du processus de préservation dans son ensemble suit les mêmes exigences d'excellence.

## 3. LA QUALITE ORGANISATIONNELLE

La mise en place d'une démarche qualité au niveau de l'organisation même d'un service dédié à la préservation numérique est un projet global mobilisant l'ensemble des acteurs qui y sont impliqués. Trois thématiques applicables peuvent y être identifiées.

### 3.1. Les processus métier

Le premier axe d'une démarche qualité consiste à inventorier et documenter les différents processus métier mis en œuvre pour la préservation à long terme de documents numériques. L'intérêt d'une telle initiative est multiple :

- Tout d'abord, la formalisation des processus métier constitue la colonne vertébrale de la documentation du service. Dans la perspective d'une conservation de documents à très long terme, les

informations relatives à toutes les procédures qui leur sont appliquées ont un sens évident, et constituent la base de la mémoire du service. Cette notion de transmission de la connaissance métier est particulièrement importante lorsque la rotation ou la mobilité des acteurs est importante (mutations, départs à la retraite, contrats de travail à durée limitée, etc.), et sa bonne gestion est primordiale ;

- Ensuite, cette formalisation permet une auto-évaluation permanente de l'organisation, par la revue et la critique régulière, et donc une amélioration continue des performances globales du système et du service mis en œuvre ;

- Dans une même perspective d'évaluation et d'amélioration, elle permet une comparaison plus efficace du mode de fonctionnement de deux structures qui auraient une démarche similaire de documentation de leurs processus métiers ;

- Enfin, la formalisation des processus métiers est un excellent support pour l'amélioration de la communication interne ou externe : elle constitue un moyen de lutte contre le cloisonnement interne interservices, ou permet une plus grande transparence dans les méthodes de travail, ce qui est un point essentiel pour établir et maintenir des relations de confiance avec les partenaires, les tutelles ou les utilisateurs.

Bien entendu, les processus ne sont jamais tout à fait figés et évoluent dans le temps ; le référentiel doit être mis à jour en fonction des changements apportés.

Le CINES a initié cette formalisation il y a près de deux ans, les résultats sont disponibles en ligne pour une plus grande transparence vis-à-vis de ses utilisateurs.

### 3.2. La gestion des risques

La méthodologie de pilotage de projet par la gestion des risques est issue du monde industriel, et est aujourd'hui très répandue dans les secteurs du management et des services.

La préservation des documents numériques est une action principalement préventive, d'où la nécessité d'intervenir avant que les dégradations se produisent. C'est un ensemble de compromis à la fois entre les impératifs immédiats des producteurs d'information et les besoins à long terme des utilisateurs, et entre les besoins et les moyens mis en place.

La préservation des documents numériques est un projet comme un autre, et comme toute autre activité, elle génère des risques. L'objectif n'est pas ici de les supprimer, mais bien de déterminer le niveau de risque acceptable. Une telle méthode est bien définie et a déjà fait ses preuves ; le plan de gestion des risques, qui résulte de cette démarche, est en place dans le service d'archivage du CINES, et est révisé deux fois par an.

### 3.3. La certification

La certification est l'aboutissement de la consolidation d'un organisme et/ou d'un service. Elle

matérialise la reconnaissance de la qualité et du professionnalisme, donc un moyen d'instaurer des relations de confiance avec les communautés d'utilisateurs, et peut constituer, en outre, un levier pour obtenir des budgets auprès des organismes de tutelle.

Plusieurs types de certification sont envisageables :

- Les certifications généralistes, telles que ISO 27000 (sécurité informatique), ISO 9000 (qualité), CMMI (ingénierie), ITIL (services), etc. ;

- Les certifications spécifiques à l'archivage pérenne, parmi lesquelles des accréditations : DSA (bonnes pratiques), DRAMBORA (gestion des risques), TRAC (liste de critères) et des certifications en cours d'élaboration : AFNOR (Z42-013) avec le SIAF, ISO 16363 (European Audit Framework) avec le CCSDS.

La démarche de certification est un projet conséquent, tant sur le plan de l'investissement humain qu'elle requiert pour la conduite du projet ou le pilotage des changements requis, que sur le plan de l'investissement financier puisque des audits externes réguliers sont à prévoir. Il convient par conséquent de bien identifier le type de certification qui aura le plus d'impact à la fois sur la communauté d'utilisateurs et sur les organismes de tutelle.

Le CINES a obtenu un agrément du SIAF (Service Interministériel des Archives de France) pour la conservation d'archives publiques intermédiaires. Sa plateforme d'archivage électronique a reçu, en mars 2011, l'accréditation DSA (Data Seal of Approval), qui est attribuée par un comité d'experts internationaux aux centres de préservation numérique ayant mis en place des procédures d'assurance qualité afin de garantir l'accessibilité et l'intelligibilité des informations qui leur sont confiées. Parallèlement, le CINES teste actuellement, dans le cadre d'une initiative européenne, le projet de certification ISO 16363 destiné spécifiquement aux services d'archivage numérique pérenne.

## 4. CONCLUSION

De la qualité des objets numériques dépend grandement la facilité de leur préservation dans le temps ; il ne faut toutefois pas réduire cette qualité au seul aspect technique des objets numériques, tant la qualité des processus de conservation qui leur sont appliqués est toute aussi cruciale.

Une démarche qualité pour la conservation à long terme représente un investissement conséquent et immédiat dont les effets ne seront perceptibles qu'à long terme ; de plus, une telle initiative, et son aboutissement qu'est la certification, requiert l'adhésion de toutes les ressources humaines impliquées dans le processus de conservation.

Pourtant, les indicateurs et les métriques sur la qualité ne sont pas encore clairement définis. Alors, rendez-vous dans trente ans ?