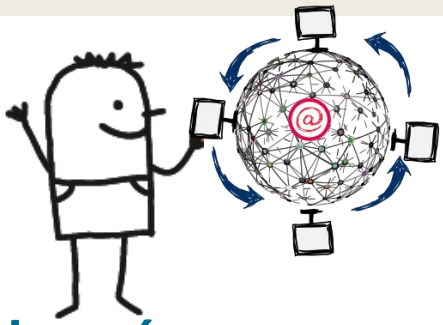
LES DÉFIS DU NUMÉRIQUE

Comment être sûr que ce fichier
n'a pas été modifié ?

Pourrons-nous relire nos données
dans dix ans?

Je croyais avoir détruit ce mail
confidentiel et il est encore là :
comment est-ce possible?

Défis techniques, défis humains



Intégrité et pérennité

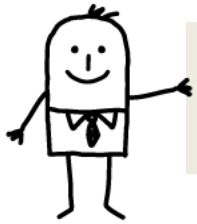
Le numérique présente des caractéristiques qui constituent une double rupture avec les pratiques d'écriture, de communication, de stockage et de recherche d'information que l'on connaissait avant avec le support papier :

- un fichier numérique est facilement modifiable: **intégrité!**
- les supports ne se conservent pas « tous seuls » dans la durée: **pérennité!**

La maîtrise des données

Le numérique présente un autre défi : celui de la maîtrise des données. Cette question comporte deux aspects :

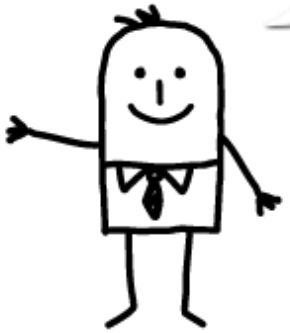
- celui de la constitution et de l'exploitation d'**immenses bases de données**, ce qu'on appelle les données structurées, et
- celui de la production et de l'exploitation des **milliards d'informations non structurées** créées chaque jour au travers de messages électroniques, de posts sur Internet, d'images voire de sons de toutes sortes. Ce qui n'est pas sans conséquences juridiques…



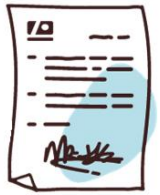
Les supports numériques sont porteurs de défis techniques.
L'explosion des données dans les réseaux est porteuse de défis humains.

L'INTÉGRITÉ

L'intégrité est la caractéristique d'un document ou d'un objet d'information, quelle que soit l'étape de son cycle de vie, dont le contenu d'information est identique à ce qu'il était à sa création.



Intégrité



1

Comment être sûr que ce fichier n'a pas été modifié ?

Stricto sensu, **l'intégrité du support** exige que celui-ci soit préservé en l'état. **L'intégrité d'un document**, constitué d'une information sur un support (toujours vrai à l'ère numérique), exige que le message n'ait pas été altéré depuis sa création et sa validation, même si le support proprement dit peut avoir subi des modifications, dès lors que les modifications n'affectent pas l'écrit.

Ainsi un contrat manuscrit dont la marge aurait été déchirée pourrait toujours être considéré comme un document intègre, alors qu'un contrat papier dont une des signatures aurait grignotée par une souris ne le serait pas.
De même, un fichier numérique qui a été recopié sur un autre support ou converti dans un autre format est intègre si le message porté sur le support est bien identique au message initialement produit par l'auteur, dans tous ses aspects informatifs, dès lors que l'on peut constater, mesurer et tracer les modifications apportées au support sans impact sur le contenu.

C'est la traçabilité des modifications ou tentatives de modification qui permet de dire qu'il n'y a pas eu de modification de nature à entamer l'intégrité du document.

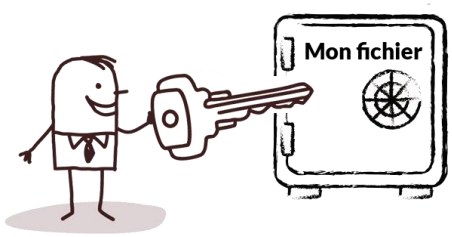
Intégrité

2

Les technologies qui garantissent l'intégrité sont l'empreinte, le chiffrement et l'horodatage

L'empreinte électronique est le résultat d'une fonction mathématique de hachage effectuée sur un fichier ; on obtient une suite de caractères extraits de façon unique d'un fichier donné ; si le fichier n'est pas modifié, la même fonction donnera la même empreinte autant de fois qu'elle sera utilisée ; si le fichier est modifié, même d'une simple virgule, l'empreinte sera différente. L'empreinte électronique permet donc de vérifier la non-modification d'un fichier numérique, donc son intégrité technique.

Un autre type d'algorithme permet de **chiffrer** le fichier ou simplement l'empreinte du fichier, à l'aide d'un jeu de clés asymétriques (une clé privée et une clé publique). C'est ce qu'on appelle communément **la signature électronique ou le scellement**.



La technique **d'horodatage** permet d'ajouter une date certaine à l'empreinte ou à la signature. Il faut rappeler qu'**il n'y a pas de preuve possible sans date certaine.** La norme internationale ISO 8601 spécifie une représentation de la date et de l'heure universelle et la RFC 31613 (recommandations pour Internet) définit un protocole d'horodatage.

Intégrité

3

Précision à propos de la signature électronique

La signature électronique n'est pas l'équivalent de la signature manuscrite au bas d'un contrat qui indique que le signataire est d'accord avec le contenu. Il s'agit plutôt d'un cachet apposé sur l'enveloppe pour dire que le contenu n'a pas été ouvert. Avec le courrier papier, c'est la Poste, avec son autorité et ses procédures, qui garantit l'intégrité du courrier. Dans l'environnement numérique, ce rôle, qui requiert des compétences et un équipement technique élaboré, est joué par des entreprises appelées « tiers de confiance », dont la Poste du reste.

À noter que la **même technologie** est utilisée à la fois pour apposer son **consentement** à un document numérique (le déchiffrement avec la clé publique sert à authentifier le signataire du document) et pour garantir la non-modification d'un document quelconque lors d'une transmission en identifiant l'auteur de la transmission, par exemple un tiers. Le détenteur de la clé privé doit posséder un certificat nominatif délivré par une autorité de certification.

Et il ne faut pas oublier qu'avant d'assurer l'intégrité d'un acte, il faut s'assurer de son authenticité, c'est-à-dire vérifier l'identité de l'auteur et la date du consentement manifesté par la signature de l'auteur.

Intégrité

4

Le point de départ du contrôle d'intégrité

La notion d'intégrité ne signifie rien si l'objet contrôlé n'est pas pris en charge depuis sa création.

L'intégrité d'un document archivé n'a de valeur que si on a constaté au départ, à la création du document, que celui-ci est complet et exact, Que vaudrait un document soigneusement déposé dans un coffre-fort et dont l'intégrité est totalement préservée depuis son entrée dans le coffre-fort si ce document est un faux ?

L'intégrité est indissociable de traçabilité

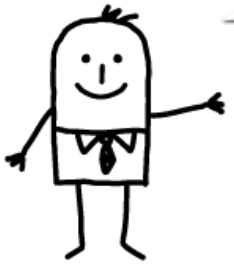
Outre l'intégrité du support et des données, il faut donc tracer la vie de ce document (support + données) depuis son achèvement, **depuis l'acquisition de son statut de document engageant ou à risque**.

Il convient d'enregistrer les événements survenus au cours du cycle de vie d'un document archivé : transfert de système, ajout de métadonnées, mais aussi le fait qu'il ne lui est rien arrivé de dommageable (pas d'accès non autorisé, pas de nouvelles données dans le corps du document, etc.).


LA PÉRENNITÉ

La pérennité d'un document ou plus précisément celle d'un support est sa **capacité à « passer les années »** (*per annos* en latin). On parle aussi de durabilité des supports.

Et **la pérennisation** est la démarche qui permet de donner aux documents la capacité de se conserver plus longtemps.

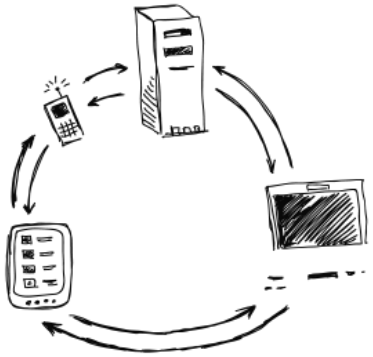


Pérennité

1

Contrairement au papier qui « s'auto-conserve » plus ou moins pendant des décennies, le support numérique est très volatil.
On parle de l'obsolescence des supports numériques qui ne sont pas capables de se maintenir au-delà de quelques années.

Le support numérique n'est pas un objet homogène.
Il faut distinguer trois composantes complémentaires :



1. **le support d'enregistrement des données,**
2. **le format d'encodage des données,**
3. **le logiciel de lecture.**

Les trois éléments sont indispensables pour pouvoir réutiliser correctement un document. La conservation doit donc les prendre en compte tous les trois.

Pérennité

2

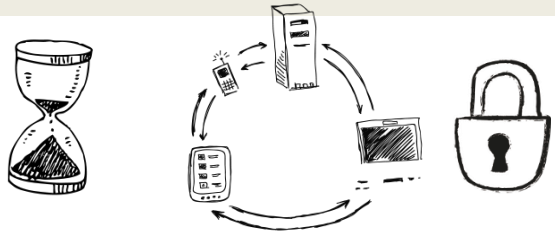
Support, format, outil de lecture

- **le support d'enregistrement des données** est le matériel de stockage proprement dit, c'est-à-dire les bandes ou les disques sur lesquels les fichiers numériques sont stockés ; leur durée de vie ,selon la qualité, peut aller jusqu'à une quinzaine d'années, ce qui est insuffisant pour la conservation à long terme ;
- **le format d'encodage des données** est le code qui permet de traduire chaque mot, chaque point d'une photo, chaque signal, en suites de 0 et de 1 (bits) pour être enregistrés sur le disque ou la bande, et surtout de les retraduire dans l'autre sens, en permettant aux 0 et aux 1 de redevenir signifiants ;
- **le logiciel de lecture** est l'outil qui « donne à voir » les mots, points et signes restitués à l'utilisateur ; la durée de vie de ces logiciels est liée à l'évolution technologique et au marché ; si un outil n'a plus suffisamment d'utilisateurs, il ne sera pas maintenu.

L'outil de lecture est lié au format. Si le code du format n'est pas public (ouvert), le lecteur est dépendant du propriétaire (éditeur) de ce format.
Le support peut être géré directement par l'entreprise, l'institution ou l'individu (serveur, disque externe) ou confié à un prestataire spécialisé: c'est le cas des tiers-archiveurs et c'est le cas du cloud.

Avec le cloud, les fichiers de données sont stockés quelque part (dans un « nuage »), accessibles par des identifiants via le réseau mondial.

Pérennité

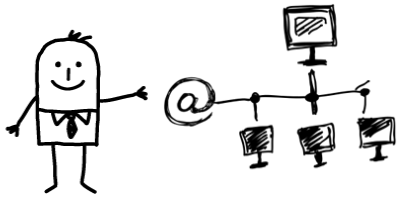


3

Compte tenu de l'obsolescence des technologies, la question de la pérennité (format, support, outil de lecture) doit être posée dès que le document doit être conservé plus de 5 ou 10 ans.

Pour l'archivage, il faut attirer l'attention sur le fait que certaines bases de données, avec des données datant de plusieurs années ou dizaines d'années sont consultées régulièrement ; lorsque le logiciel de gestion de cette base évolue, les données sont migrées sur la nouvelle version de l'outil. C'est le cas par exemple avec la base des patients dans un hôpital dont les données remontent à plusieurs décennies; c'est le cas également des archives audiovisuelles en ligne sur le site de l'INA.

Mais lorsque les documents numériques ne sont pas consultés régulièrement et sont enregistrés sur des supports externes, l'obsolescence est plus difficile à maîtriser. Comme le dit un slogan du groupe de recherche international InterPARES :



« Conserver un document électronique est à proprement parler impossible ; seule la capacité de le reproduire peut être préservée ».

Cela signifie que le document numérique n'est pas un objet fixe mais une suite de données caractérisées qui permettent de reconstituer le document avec fiabilité (qui a dit quoi quand).

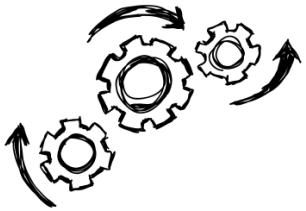
Pérennité

4

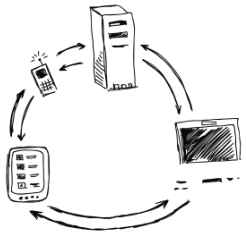
Il existe des solutions technologiques pour répondre aux enjeux de pérennité des documents à conserver dans le temps.

Deux grandes approches s'opposent:

1. Préserver l'environnement de production du document, c'est-à-dire le logiciel initial, et le réactiver pour lire le document: c'est ce qu'on appelle **l'émulation**;
2. Transférer les données dans un nouveau format et/ou vers un nouveau support plus récent et donc plus efficient: c'est ce qu'on appelle **la migration**.

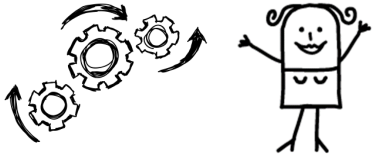


Ces solutions vont évoluer grâce aux recherches et expérimentations en cours dans les secteurs d'activité les plus concernés (industrie automobile ou aéronautique) ou dans les associations professionnelles [*ce sujet sera développé lors de la semaine 6*].



Dans l'entreprise, c'est le principe de la migration qui prévaut. Les opérations de **migration** sont plus facilement réalisées par des **prestataires spécialisés** qu'en interne, compte tenu de la spécificité et de la technicité des opérations de migration, de même qu'une entreprise confiera la restauration de ses documents papier, endommagés par une inondation ou un accident matériel, à des prestataires spécialisés, au fait des meilleures techniques.

Pérennisation

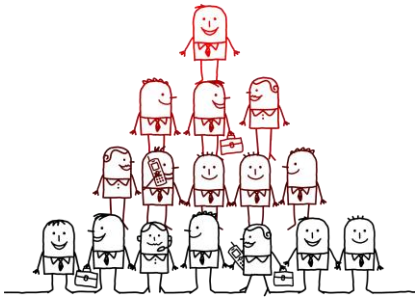


La pérennité est la capacité intrinsèque du document ou du fichier à s'auto-conserver.
Si la pérennité initiale est faible par rapport à la durée de conservation nécessaire, il faut la renforcer par des mesures organisationnelles et techniques. C'est le rôle de la pérennisation.

Une démarche globale de pérennisation comporte deux grands types d'actions:

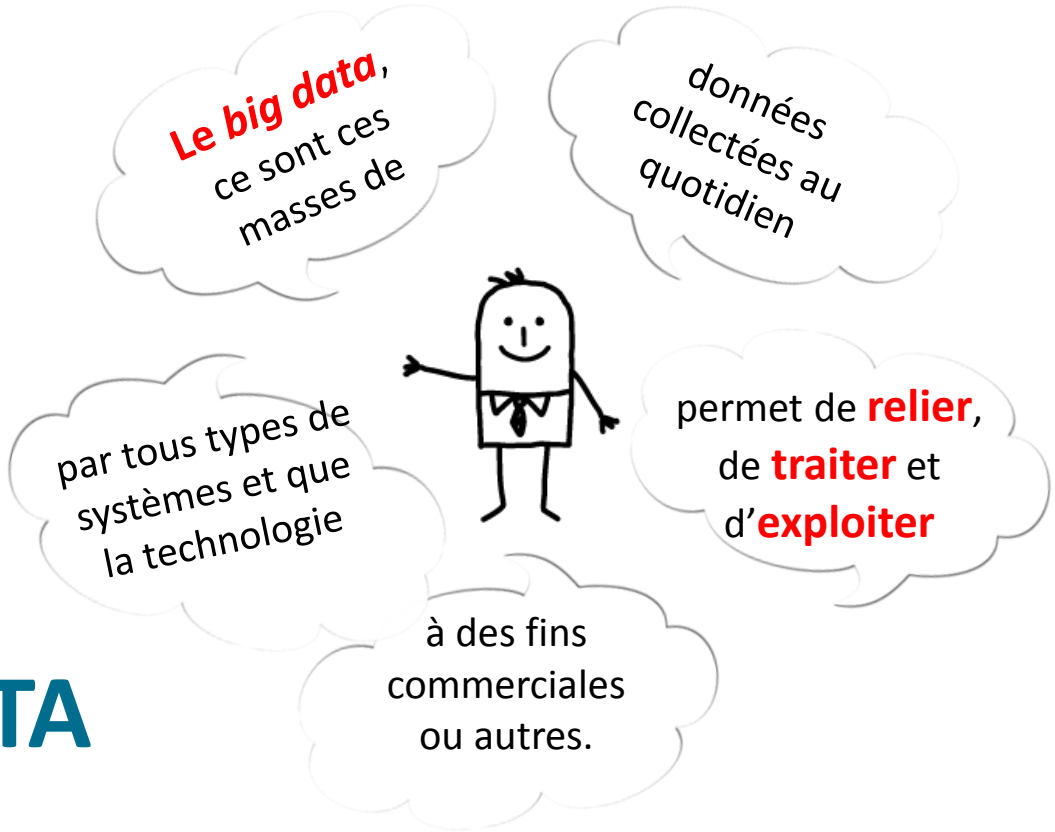
1. Avant l'archivage, il faut se préoccuper de **créer les documents dans le format le plus approprié** en s'appuyant sur les normes en vigueur sur les formats ouverts (PDF, XML…) et le plus « pérenne » ; corriger les défauts de pérennité et d'intégrité d'un document numérique après coup est toujours plus lourd et plus cher;
2. Après l'archivage, il faut veiller à la qualité des données pour **opérer les migrations de manière préventive** plutôt que curative; les données initiales, avec leur signification doivent être migrées d'un support à l'autre et d'un format à l'autre, tout en garantissant que le document d'arrivée est identique au document de départ dans sa valeur de preuve ou de mémoire.

La démarche de pérennisation inclut d'effectuer et de faire effectuer une veille sur les nouvelles technologies de pérennisation.



EN MATIÈRE DE MAÎTRISE DES DONNÉES ET D'ARCHIVAGE, L'AMONT (LA CRÉATION) CONDITIONNE L'AVAL (LA CONSERVATION).

LE BIG DATA


Le *big data*, ce sont ces masses de
données collectées au quotidien
par tous types de systèmes et que la technologie
permet de **relier**, de **traiter** et d'**exploiter**
à des fins commerciales ou autres.

Big data: masse et pertinence

La question de la conservation des documents archivés n'est pas la seule question posée par l'archivage dans l'environnement numérique.

En amont des questions techniques de l'archivage numérique se pose la question de la pertinence de ce qui est archivé.

Les chiffres donnés par les experts sur l'explosion des volumes de données sont vertigineux, même si la capacité croissante des supports de stockage masque en partie le phénomène.

On parle aujourd'hui du phénomène « big data » qui désigne les masses de données produites de manière continue par les outils de communication et d'échanges d'information, ou diffusées sur les réseaux, quel que soit le producteur.

Ces données sont si volumineuses et si dispersées que les moyens classiques d'exploitation des entrepôts de données ne suffisent plus à en venir à bout.

L'exploitation de ces masses de données est possible grâce à de nouveaux algorithmes et de nouveaux traitements ; c'est un segment croissant de l'économie numérique.

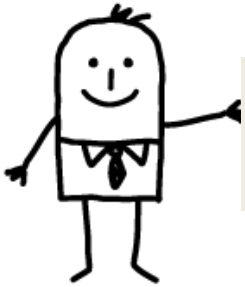
Toutes les données sont concernées

Les données structurées

- **Des systèmes** de plus en plus sophistiqués, connectées ou connectables
- **Traçant automatiquement les moindres gestes des individus**, avec une valeur souvent éphémère de transaction, de sécurité ou d'échange.

Les données non structurées

- **Des informations** (mails, SMS, posts, commentaires, transmissions, photos, …)
- Envoyées volontairement sur le réseau par chaque individu
- **Sans contrôle de leur devenir par leur émetteur**, sans règle de vie pour leur stockage, leur accès, leur destruction, sans conscience des risques associés.



Les données, structurées ou non, sont exploitables, à des fins distinctes de leur finalité d'origine.

Comment contrôler l'accumulation exponentielle des données ?

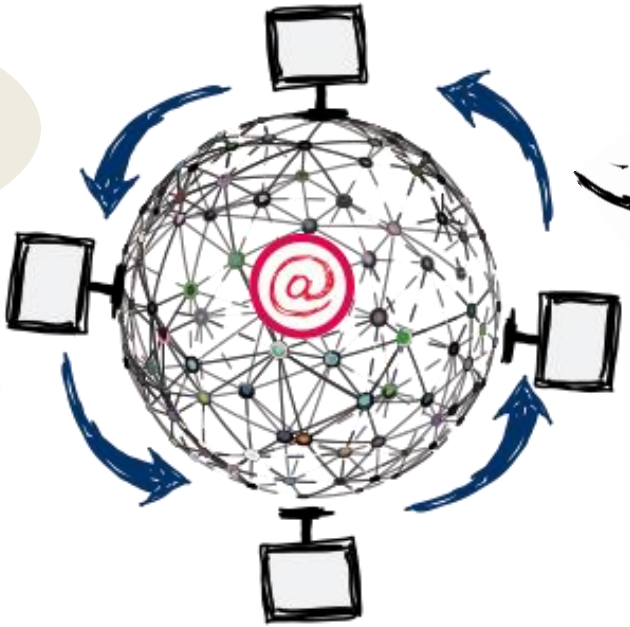
Le mieux est la **prévention**: il s'agit de **contrôler les flux** en intégrant des règles de vie (durée de conservation, accès) au document dès la création ou la réception de l'information, c'est-à-dire ***a priori*** puis en vérifiant l'application des règles.

Oui, mais si les règles n'ont pas été correctement définies, il faudra organiser **des mesures curatives** pour réduire les risques en traitant **le stock *a posteriori.***

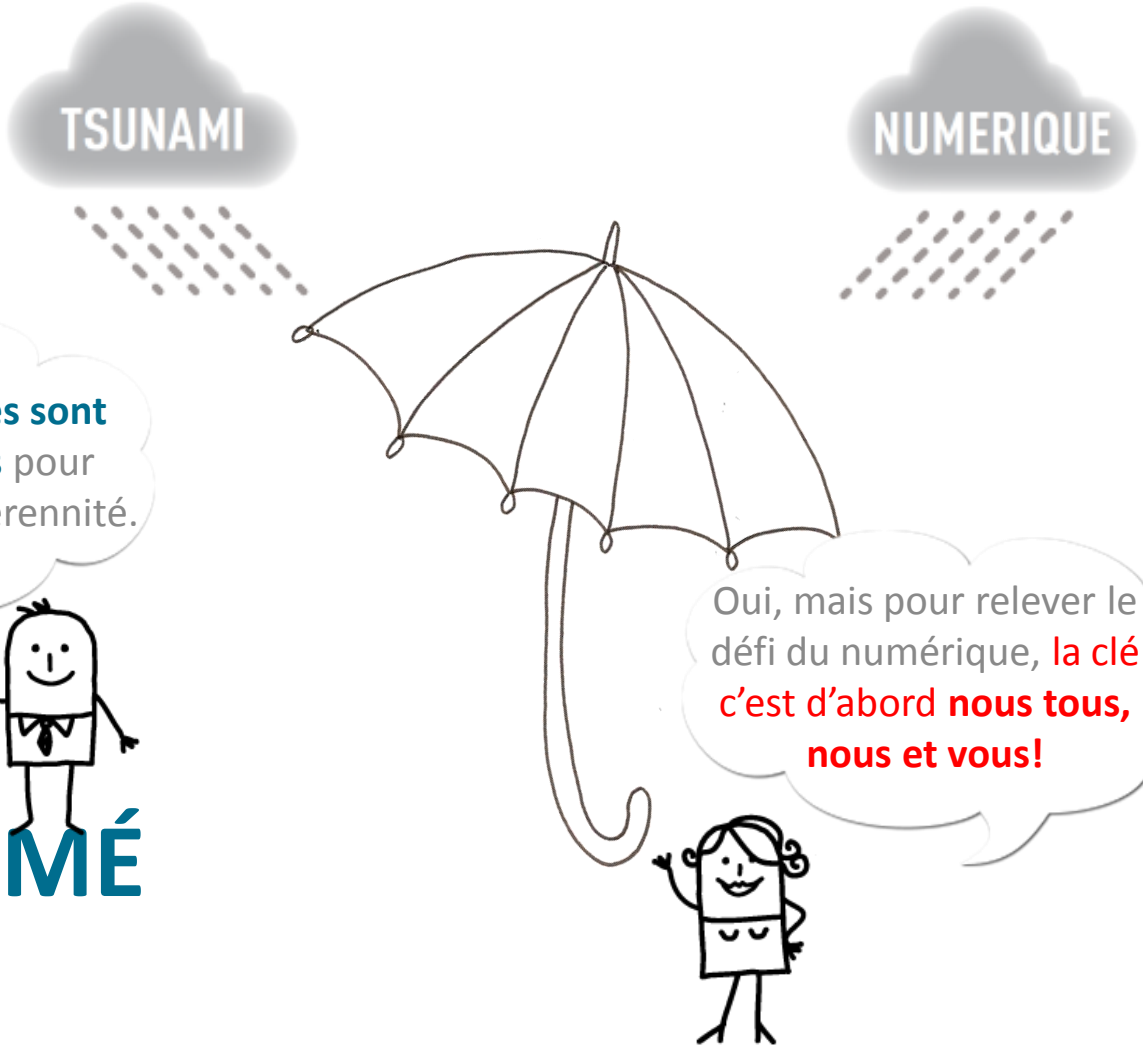
Défi humain: réponse managériale!

Face à la production vertigineuse de documents et de traces, à caractère éphémère et souvent personnel, la réponse est d'abord d'ordre managérial: il faut contrôler les flux d'information (émission et réception) en édictant des règles et contrôler l'application de ces règles.

Contrôler
les flux et donner
des règles



Contrôler
l'application des
règles et corriger


TSUNAMI
NUMERIQUE
Les technologies sont **indispensables** pour l'intégrité et la pérennité.
Oui, mais pour relever le défi du numérique, la clé c'est d'abord **nous tous, nous et vous!**

EN RÉSUMÉ