Les stratégies de préservation

Bruno Bachimont, Université de technologie de Compiège, France

Prologue

Préserver des contenus numériques

- Les constituants d'un document numérique sont
  - Le contenu du document
  - Les métadonnées décrivant le document.
- Le contenu est donné comme une séquence ou flux de caractères pris dans un alphabet fini de référence A.
- Les métadonnées
  - Elles identifient cet alphabet utilisé pour encoder le contenu du document ainsi que d'autres informations caractérisant le document comme objet (son auteur, son identifiant, etc.).
  - Elles sont codées comme des séquences de caractères pris dans un alphabet de référence B. Les alphabets A et B peuvent être ou ne pas être identiques.
  - La structure des métadonnées et la nature de l'alphabet B sont définies et spécifiées par l'archive qui gère et préserve le document numérique.

Formats

- Formats
  - Un format est dit binaire (*binary data format*) quand les contenus sont stockés comme des suites de zéro et de un.
  - Un format est dit textuel (*text-based data format*) quand les contenus consistent dans des caractères issus d'un alphabet donné (par exemple ASCII ou Unicode).
- Un document stocké à l'aide d'un format textuel voit finalement les caractères qui le codent traduits dans la séquence binaire qui leur correspond selon l'alphabet de référence. Au final, c'est bien un séquence de 0 et de 1 qui sont stockés.

Formats usuels

catégorie	Types de données	Formats standards
Données	Alphanumériques	PDF, PostScript, ASCII, SQL
Texte structuré	Alphanumérique, référence à des images, balises	PostScript, PDF, TeX, XML, HTML, XSL
Document de conception	Bitmap, vectoriel, alphanumérique	HPGL, PostScript, EPS, CGM, TIFF, ASCII
Bureautiques	Alphanumérique, vectoriel, bitmap, graphique animé	PDF, PostScript, XSL, RTF, ASCII, XML, TIFF, CGM
Présentations	Bitmap, vectoriel, alpanumérique, animation	PostScript, PDF
Image	Bitmap	PS, PDF, TIFF, GIF, JPEG
Audio	Audio	MPEG-1 (1-2-3), MP3, MIDI
Vidéo	Vidéo	MPEG-1, MPEG-2, MPEG-4
Données géographiques	Bitmap, vectoriel, alphanumérique	PS, EPS, HPGL, TIFF, ASCII, CGM
Multimédia interactif	Tout !	MPEG-1, MPEG-2, SMIL

On distinguera donc :

- Format de fichier / format d'écriture :
  - Le format qui organise la manière d'écrire une information sur un support d'enregistrement
  - Format DVD, LTO, etc.
- Format de codage :
  - Le format qui code un contenu :
    - Format textuel
    - Format binaire

UNE DISTINCTION

Refaire, ré-exploiter ?

Deux objectifs bien distincts

- Préserver pour refaire
- Préserver pour ré-exploiter

Refaire...

- Principe :
  - reproduire un contenu / œuvre dans son état d'origine ou équivalent à cette origine.
- Enjeu :
  - Une conformité expressive;
  - Une fidélité sémantique.
- Exemple:
  - Montrer une photo, une vidéo.

Ré-exploiter…

- Principe :
  - Accéder à l'information contenue dans une ressource ou un document;
  - Il faut pouvoir utiliser l'information sans avoir forcément besoin de la montrer.
- Enjeu:
  - Assurer la fidélité sémantique;
  - Pas de conformité expressive.
- Exemple :
  - Données scientifiques : refaire une simulation à partir de données anciennes.

Deux types d'objets numériques

- Données :
  - Ensemble de bits représentant une information sur quelque chose.
- Contenus :
  - Ensemble de bits codant une information nécessaire pour construire un objet perceptible.

Données : bits, décodage, interprétation

\+ Schéma de décodage = Valeurs

Valeurs + Schéma de d'interprétation = Données

Des bits aux données

- Le schéma de décodage permet de passer des bits comme des valeurs : nombres, caractères, etc.
- Le schéma d'interprétation permet de passer de ces valeurs à des données des données :
  - on interprète les valeurs comme des nombres qui sont des mesures métriques.
- Les données sont indépendantes de toute publication/visualisation qui peut donc être arbitraire par rapport aux données.
  - Ce qui compte: le schéma d'interprétation qui permet de passer de la trace (les bits) à des données abstraites, indépendamment de comment on les visualise : par exemple, la donnée 2 mètres, que ce soit en russe, en chinois, par écrit ou sur écran, etc.

Contenus: bits, décodage, visualisation

\+ Schéma de décodage = Valeurs

Valeurs + Schéma de visualisation =

Contenu

Des bits à la visualisation

- Le schéma de décodage permet de reconstruire les valeurs à partir des bits, ces valeurs étant des codes/instructions pour la visualisation
- le schéma de visualisation permet de passer de ces valeurs à un objet perceptible concret.
- L'utilisation et l'exploitation commencent à partir de la forme publiée qui dépend de l'objet concret particulier, et non de l'abstraction qu'il représente.

Les contenus ne sont pas des données

- Contenus:
  - Il n'est pas possible d'abstraire les contenus dans une représentation formelle indépendante de la forme publiée et perceptible.
    - E.g. un document ne peut être remplacé par sa représentation RDF.
  - L'enjeu est de publier les bits enregistrée pour permettre et démarrer leur utilisation et interprétation.
- Données:
  - Possible de les abstraire en une représentation formelle.
  - L'enjeu n'est pas de publier les bits enregistrés (pour les rendre visibles) mais de les exploiter.

Mais pour les contenus culturels

- Ce qui est la bonne ou l'authentique publication peut rester vague ou sous-déterminée.
- Quand les données sont complexes, ou sont perdues, ou encore bas niveau, il faut revenir à l'intention de l'auteur ou à la documentation sur la sémantique de l'œuvre.
- Mais alors, cela devient de l'herméneutique, et non de la formalisation.

- Les données se formalisent, les contenus s'interprètent

Les stratégies de préservation

Préservation traditionnelle ou passive

- Deux principes :
  - Conservation Préventive
  - Restauration active
- Les stratégies répondent aux causes de détérioration
  - Prévention
    - Identifier les causes, voir pour chaque cause comment prévenir
    - Rappel : en contrôlant température et humidité, on va déjà loin
  - Restauration
    - Préservation matérielle: causes physiques, chimiques, biologiques

Préservation numérique ou active

- Deux principes :
  - Intervenir sur le contenant quand il est encore en bon état
    - Corruption des supports
  - Intervenir sur le contenu quand il est encore lisible
    - Obsolescence des formats
- Rapport différent au contenu:
  - On n'essaie plus de préserver les supports physiques, seulement l'information

Deux philosophies différentes

- Préservation passive:
  - Tant que ça va, on ne fait rien ;

- Préservation active :
  - Pendant que ça va encore, on fait ce qu'il faut.

Quelques problèmes avec le numérique

- Supports physiques non pérennes
- Obsolescence technologique
- Volatilité des logiciels
- Variété des versions
- Standards « orphelins »
- Normes en évolution constante
- Nouveautés sans cesse (exemple récent : iPad et concurrents)

Préservation active

- On ne peut pas se contenter de stocker les fichiers (preserve the bits)
- Il faut être proactif, sinon c'est certain que les documents numériques seront perdus
- La notion des 3 âges en archivistique (actif, semi-actif, permanent) remise en question

Stratégies de préservation

- Les différentes stratégies possibles
  - Le musée :
    - conserver les contenus, les machines et les programmes tels quels;
  - La migration :
    - faire évoluer les contenus pour que leur format technique et logiciel soit compatible avec les standards et machines du moment;
  - L'émulation :
    - Conserver les contenus tels quels, mais simuler les machines anciennes permettant d'exploiter les contenus sur des machines actuelles.
  - La description :
    - Décrire les contenus et leur exploitation aussi précisément que possible pour être capable de les reproduire et de les refaire.
  - La distribution :
    - Confier de multiples copies au « réseau » pour multiplier la probabilité de conserver une copie pour l'avenir.

Musée

Le musée

- Conserver les objets et leur dispositif de lecture d'origine
  - Permet de restituer l'environnement et les sensations originales liés aux contenus et à leur consultation
  - Ne peut être mis en œuvre à grande échelle car il faut une maintenance particulière des dispositifs associés alors qu'ils sont obsolètes
  - Complémentaire par échantillonnage des autres stratégies de préservation.

Maintenir le matériel: le numérique

Maintenir le matériel : l'audiovisuel

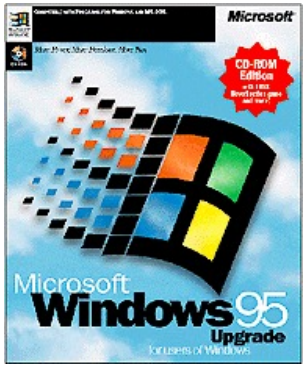
Ça dégénère…

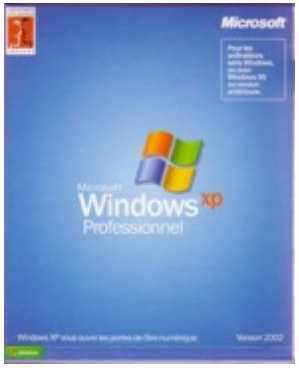
Remarques

- À court terme, une stratégie qui aide
- Mais :
  - L'équipement électronique est capricieux
    - Archiviste = technicien
  - Les machines se détériorent: où trouver les pièces ?
    - Cannibaliser d'autres machines pour en maintenir une en fonction
  - Stratégie perdante à la fin
  - Ne dit rien des contenus eux-mêmes :
    - Coupler avec une prévention / restauration des supports.

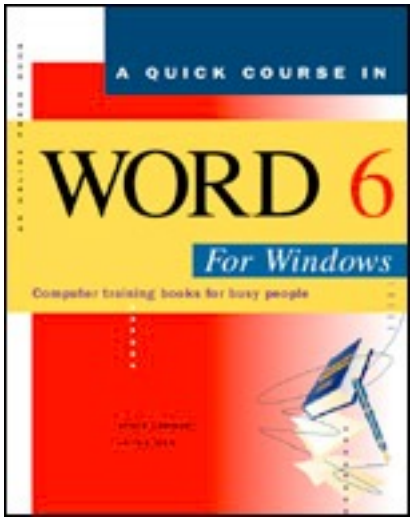
Migration

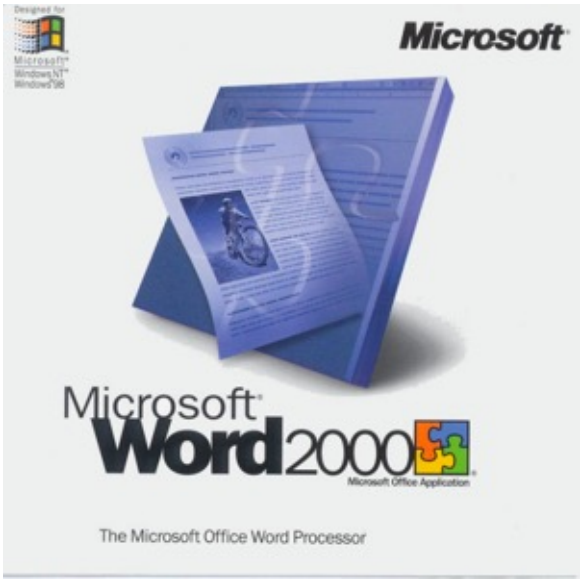
Migration de système…



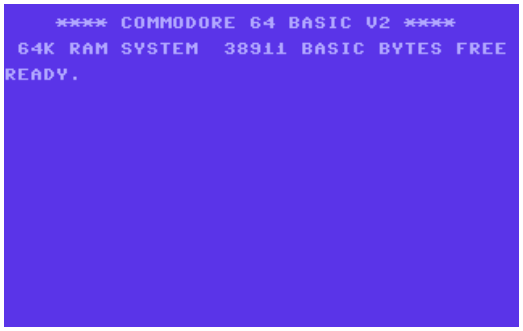


Migration d'environnement…





Migration de matériel



Migration

- Une définition
  - Migration is the periodic transfer of digital material from on hardware/software configuration to another or from one generation of computer technology to a subsequent generation. The purpose of migration is to preserve the integrity of digital objects and to retain the ability for clients to retrieve, display, and otherwise use them in the face of constantly changing technology.
  - Task Force on Archiving of Digital Information (TFADI 1996)

Formats de migration

- La question est de savoir dans quel format migrer:
  - Choisir des formats publics, avec un accès complet à leur syntaxe et leur sémantique ;
  - Le format doit être standardisé par une institution de référence (ISO, W3C, ANSI, etc.).
  - Le format doit être bien accepté et diffusé : le marché offre alors des outils de lecture et d'affichage ;
  - Le format ne doit pas être l'objet de brevet et de license, de manière à permettre un usage libre.

Pourquoi migrer...

- Les supports numériques sont plus fragiles que les supports analogiques et classiques
- Profiter des progrès technologiques :
  - Avoir de meilleurs capacités de stockages;
  - Avoir de meilleurs formats;
  - Profiter d'outils conformes à l'état de l'art du moment;
  - S'inscrire dans les pratiques et usages du moment :
    - pas nécessaire d'avoir d'apprentissage spécial pour accéder au document
- Pallier l'obsolescence technique
  - Rareté, surcoût des matériels et de leur maintenance
  - Perte de savoir faire.

Plusieurs migrations numériques

- Rafraîchissement des supports
- Réplication des données
- Recopie analogique
- Transformation numérique.

Migration : rafraîchir les données

- Principe:
  - Migrer sur un nouveau support mais avec le même format et contenu binaire.
  - Action préventive pour éviter les pertes d'information dues aux défauts physiques.
- Pas toujours possible en fait: les supports physiques ont des formats spécifiques d'enregistrement.
- TFADI:
  - Migration includes refreshing as a means of digital preservation but differs from it in the sense that it is not always possible to make an exact digital copy or replica of a database or other information object as hardware and software change and still maintain the compatibility of the object with the new genaration of technology.

Migration: réplication numérique

- Changer de support numérique implique un changement de format:
  - K7 magnétique : organisée physiquement en séquences de blocs pour un accès séquentiel aux octets
  - DVD: séquences et blocs pour un accès direct.
- Intérêt:
  - Contenu reste le même.
- Problème:
  - Les tests d'intégrité reposent souvent sur l'organisation physique des données.
  - Ils sont falsifiés par la réplication.

Migration: vers du non numérique

- Les supports non numériques sont encore les supports les plus utilisés pour la préservation à long terme.
- Le plus utilisé : le microfilm
  - Faible détérioration
  - Bonne résistance aux dégradations.
  - Facile à lire, facile d'accès.
  - Pas adapté pour les vidéos, bases de données, etc.

Microfilm

- Utilisé pour des copies analogiques
- Utilisé pour des copies numériques : le code est archivé, et non le résultat de l'affichage qu'il permet de calculer.
  - Les métadonnées conservent la manière de lire et interpréter les données binaires archivées.

Migration : transformation

- Principe : migrer les contenus vers des formats conformes aux finalités de la préservation
  - Formats supportés par les applications courantes pour un accès facilité
  - Formats permettant de compresser et optimiser l'encombrement des données.
- Approches habituelles :
  - Upgrade dans une famille de logiciels.
    - Compatibilité descendante limitée : garder les versions anciennes du logiciel, ou migrer toutes les données.
  - Exports dans un format tiers
    - Formats d'échange qui s'imposent (XML, RTF) mais souvent limité.
- Difficultés:
  - Pertes d'information, transformations des données.

Migration : avantages

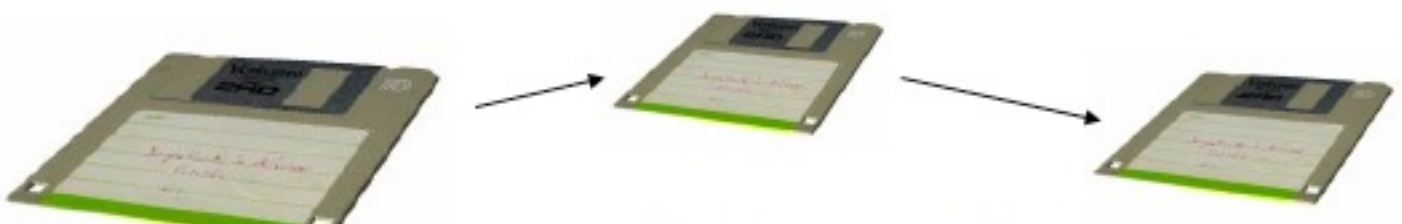
- Les documents sont toujours accessibles et bénéficient des nouvelles modalités de communication (e.g. Web pour les livres).
- Les documents restent toujours exploitables par les outils du moment.
- Les contenus bénéficient des progrès apportés par les nouveaux formats;
- La migration repose sur les savoir faire du moment : transformation des contenus par une pratique permanente, accès et exploitation.
- La migration impose un suivi permanent du fonds et le maintient vivant.

Migration: inconvénients

- La prolifération des formats implique un certain bricolage des solutions de migration.
- Les migrations successives entraînent des modifications qui altèrent l'intégrité et l'authenticité du document.
  - Documenter les migrations
  - Garder l'original

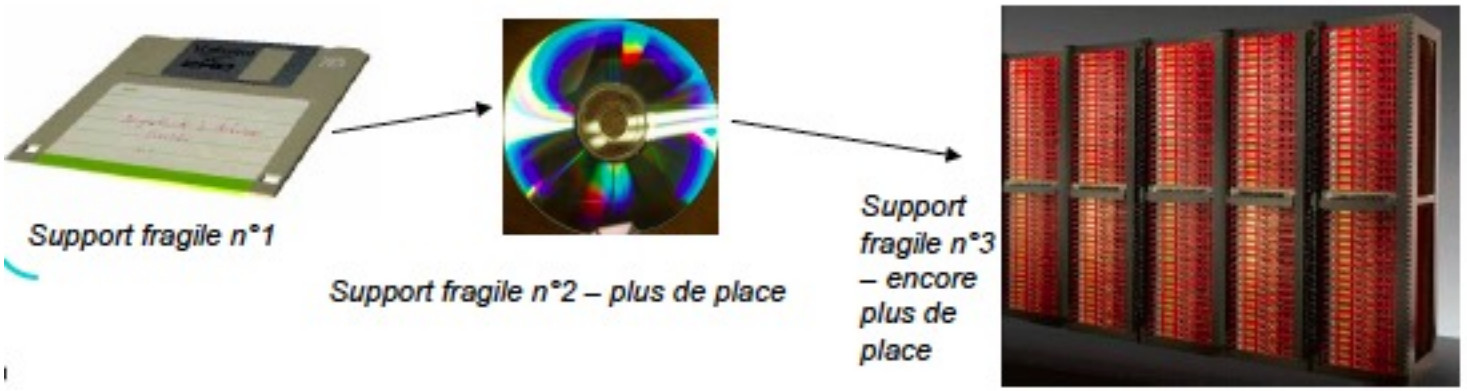
Migrations – 1

Rafraichissement des supports



Support fragile – difficile à préserver – facile à reproduire

Renouvellement de supports



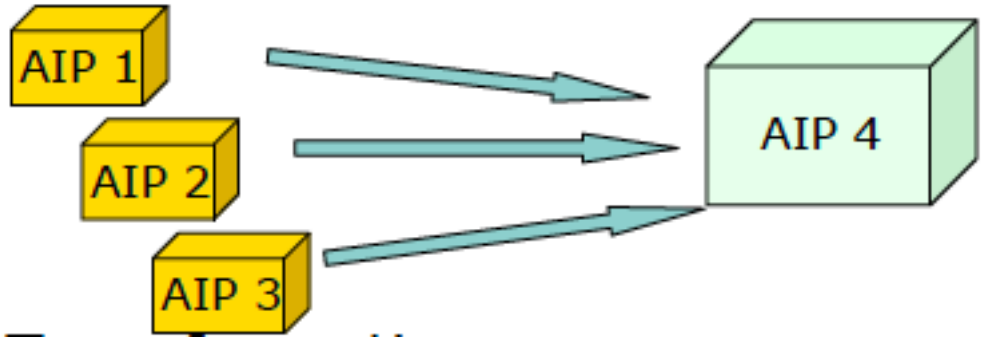
Support fragile n°1

Support fragile n°2 – plus de place

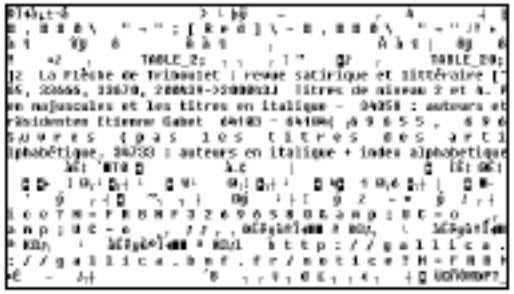
Support fragile n°3 – encore plus de place

Migrations - 2

Ré-empaquetage de l'information



Transformation de l'information



Format propriétaire – difficile à préserver

Format ouvert et libre – facile à préserver et à transformer

Émulation

Émulation

- Principe:
  - Garder les documents intacts
  - porter les outils de lecture et d'exploitation sur les plateformes du moment.
  - Migrer les outils et non les contenus !
- Approches:
  - Émulation logicielle
    - Porter l'outil de lecture
  - Émulation virtuelle
    - Porter l'outil de lecture sur une machine virtuelle, implanter la machine virtuelle sur la plateforme visée.
  - Émulation matérielle
    - Avoir une machine simulant physiquement l'ancienne.

Émulation

Environnement matériel et logiciel difficile à préserver

Environnement matériel et logiciel courant

Distinction

- Virtualisation :
  - Tout procédé rendant un document indépendant du support matériel de restitution;
- Emulation :
  - Simulation des outils de lecture de documents anciens sur des environnements technologiques modernes.
- L'émulation est un cas particulier de virtualisation

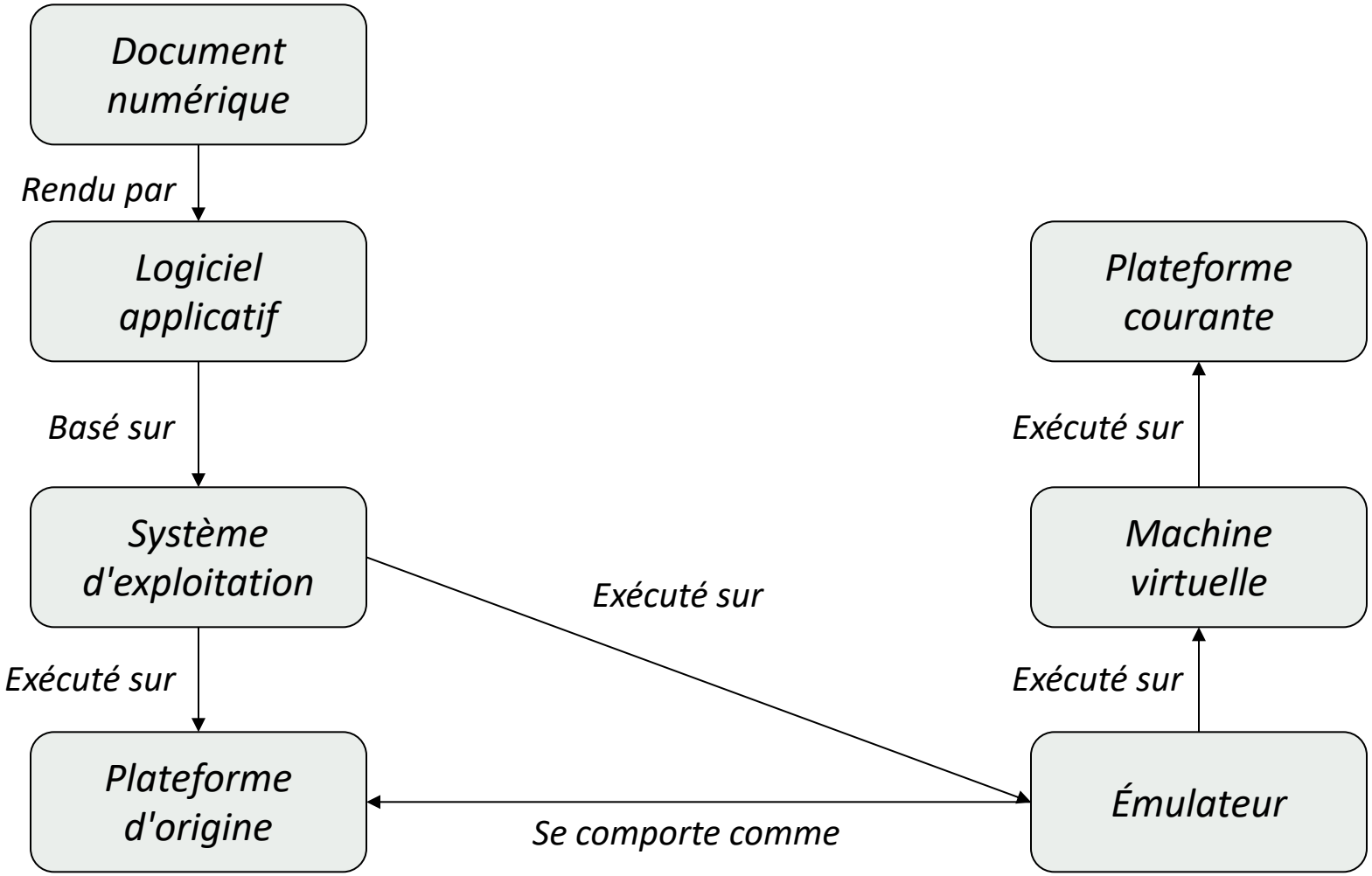
Émulation logicielle

- On s'abstient de préserver les outils de lecture. Plusieurs approches possibles :
  - Emulation d'un logiciel :
    - un émulateur est nécessaire pour chaque application, éventuellement pour chacune de ses versions.
  - Emulation d'un système d'exploitation :
    - un émulateur est réalisé une fois pour toute, et toutes les applications deviennent utilisables.
- L'émulation serait meilleure que la migration. Mais :
  - les émulateurs doivent aussi migrer, à moins de les émuler!
  - l'émulation repose sur le principe que les ordinateurs sont équivalents entre eux car ils sont tous équivalents à une machine de Turing: l'émulation parfaite est possible. Mais en réalité, les équivalences ne sont que partielles.

Émulation virtuelle

- Emulation virtuelle :
  - le traitement des différentes données et documents sont spécifiées dans le langage d'une machine virtuelle, simple, complète et non ambiguë, pouvant être implémentée sur toutes les plates formes.
  - Lorie (IBM) : universal virtual computer (UVC)
- les objets sont préservés dans leur format d'origine, mais des règles spécifient leur lecture / décodage sur l'UVC.
- chaque donnée est traitée selon un schéma logique qui lui attribue une balise sémantique.
  - Mais -1: l'UVC ne fournit qu'un ensemble réduit d''instruction et de commandes ;
  - Mais -2 : l'UVC n'est optimisée pour aucune des plates formes en particulier et les performances sont moindres.

Émulation virtuelle : Principe



Émulation matérielle

- Le principe est de recréer une ancienne machine sur une puce configurable.
- En cas d'obsolescence ou de problèmes, on peut reprogrammer un autre type de puce.

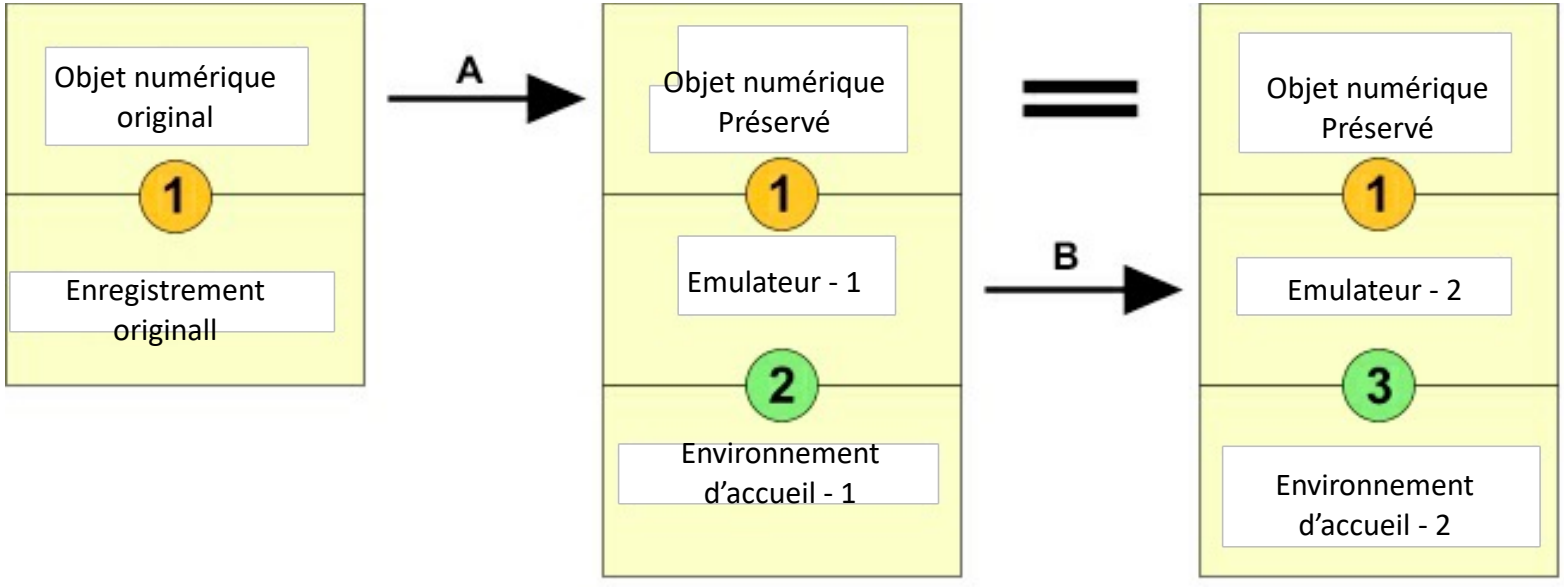
Émulation : avantages

- Authenticité garantie des contenus: on n'y touche pas !
- Tous les types de documents peuvent être archivés, y compris interactifs ou dynamiques.
- L'effort ne dépend pas du nombre de documents, mais du nombre de plateformes considérées :
  - Développer un émulateur pour chaque nouvelle plateforme
  - Porter la machine virtuelle sur chaque nouvelle plateforme.

Émulation : limites - 1

- La technique n'est pas parfaite : émuler revient à hériter des bugs et problèmes dans anciens contenus et systèmes ;
  - La combinatoire des émulations devient complexe: nouvelles versions des logiciels tous les 18 mois, conservations sur des dizaines d'années, explosion des émulations.
  - Les collections se constituent sur plusieurs années: l'émulation revient à les structurer selon leur nature technique et non selon leur structure archivistique (provenance et ordre original).

Faisabilité à long terme ?



Le temps >qui passe

Interface d'émulation

Environnement d'émulation

Émulation : limites - 2

- Utilisation des anciens logiciels: former les utilisateurs à ces anciens outils et les priver des performances des nouveaux.
  - Contradiction entre des outils et formats restés spécifiques, utilisables que dans des niches avec la nécessité de les rendre accessibles à une grande échelle (e-commerce, e-goverment, e-culture).

Rothenberg

- La seule manière de recréer les fonctions et apparences (look and feel) d'un document numérique est de le recréer dans son contexte d'origine émulé sur des machines modernes.
  - il est reste difficile de savoir quelles sont les propriétés à conserver en priorité dans une œuvre, son essence.
  - comme on est incapable d'abstraire les principales fonctionnalités à conserver, il vaut mieux garder l'objet lui-même et pouvoir le reproduire.
  - L'émulation est alors la seule solution, puisqu'elle permet de ne pas toucher à l'œuvre, contrairement à la migration, et de ne pas fausser la perspective, contrairement à la description.

Description

Principe

- Le contenu à préserver est trop fragile ou trop complexe pour être conservé dans son état d'origine :
  - Supports fragiles condamnés à court terme
  - Dispositifs complexes dont on ne maîtrise le fonctionnement que lors de la création originale.
- Il vaut mieux conserver une description permettant de récréer le contenu que le contenu lui-même.

La description ; un paradigme: la musique

- La musique (classique) repose sur une stratégie de préservation par description:
  - La partition
    - contient les instructions nécessaires pour être capable de reproduire la musique sur un instrument ; la partition n'est pas une description du contenu, mais une méthode pour le produire, un manuel d'instruction.
  - La lutherie
    - conserve les instruments et la manière de les construire. C'est le complément indispensable de la partition, et permet de retrouver la musique dans son timbre, tonalité, mélodie, etc.
  - Le conservatoire
    - Maintenir une pratique des instruments et de la lecture des partitions.
    - Transmission directe entre personnes autour des instruments et contenus.

Scoring (Rinehart)

- Le principe est de s'appuyer sur le modèle musical :
  - la partition (score) constitue une abstraction représentant l'invariant entre les différentes interprétations ou manifestations de l'œuvre et définit l'intégrité de l'œuvre.
  - Le code informatique est une sorte de partition exécutée par l'ordinateur. Mais il est trop dépendant de l'environnement d'exécution. Il faut un système de notation plus abstrait que le code, indépendant de l'environnement et aussi robuste que la notation musicale.
  - Il faut aboutir à un système de notation pour l'art reposant sur un modèle conceptuel mobilisant une ontologie ou un cadre de métadonnées.
- Mais le problème est que les modèles conceptuels ne sont pas stables, pas autant que la notation musicale (Rothenberg).

Authorial intention : variable media network

Preservation is an interpretive act

- la tension est de choisir entre figer une œuvre dans une interprétation donnée, ou de la garder vivante pour de futures interprétations.
- L'approche n'est donc pas d'avoir l'enregistrement fidèle ou authentique de l'œuvre, mais la possibilité de la recréer.
- Cette approche reprend le scoring, mais en introduisant l'intention de l'auteur.

Questionner l'auteur ?

- Un questionnaire à destination des artistes les interrogent sur leur intention artistique et leurs prescriptions pour de futures interprétations: quelles licences sont permises pour la recréation, lesquelles seraient une trahison ?
- le résultat du questionnaire donne lieu à une base de donnée, le Variable Media Kernel, multi-institutions, permettant de croiser les œuvres selon les genres et les styles.
- Approche stimulante mais repose plus sur l'auteur que sur l'œuvre: retour de l'auteur contre le structuralisme ?

Canonicalization : Clifford Lynch

- Trouver l'essence de différents formats et contenus
- Pour chaque type de données, texte, image, son, déterminer une représentation canonique du contenu: par exemple, pour le texte, son écriture en ASCII.
- Caractériser les autres types de représentation comme des spécialisations de cette représentation canonique, surchargeant de détails importants mais pouvant être oubliés: ASCII -> RTF -> Word.
- il faut donc remplir deux conditions :
  - toute représentation doit pouvoir se convertir (avec perte) dans un format canonique ;
  - les représentations multiples d'un même contenu doivent avoir la même traduction canonique.

Trouver les points saillants

- Le principe est de représenter un œuvre par des caractéristiques saillantes :
  - faire exprimer par l'auteur ce qu'il considère comme le plus important dans son œuvre ;
  - l'assister dans ce travail pour des outils qui guident l'expression de ces caractéristiques ;
  - à l'aide de ces caractéristiques, capter et conserver les parties correspondantes de l'œuvre et s'assurer que toute recréation les préserve. Ces caractéristiques et contenus correspondants sont alors les invariants de l'œuvre que la préservation doit conserver.

Typed Object Model Conversion

- Tout les contenus numériques sont des objets avec des attributs, méthodes et opérations spécifiques ainsi qu'une sémantique spécifique. Word est défini par son encodage, un mail par des attributs logiques : To, From, etc.
- les objets sont rassemblés en types, qui spécifient des valeurs ou plages de valeur pour ces attributs, méthodes.
- les types des objets spécifient les conversions des objets et les contraintes à respecter. Chaque objet consiste en effet en séquences d'octets sur lesquels on peut appliquer des conversions: mais les conversions doivent laisser invariants les attributs des objets.
- Exemple :
  - conversion de Word à PDF: les bytes sont convertis, mais les attributs textuels et visuels sont conservés.

Pierre de rosette

- un échantillon d'objets de chaque type pertinent est construit dans un format lisible.
  - on veut conserver des documents sources, disons des fichiers Words: on sélectionne des documents échantillons avec toutes les caractéristiques pertinentes.
  - on convertit ces fichiers dans un format de référence: micro-film ou même papier, car ça reste lisible indépendamment d'un système. Les fichiers obtenus constituent les documents références.
  - Etant donné un format cible, on crée dans ce format les documents références. En comparant les documents échantillons avec les documents cibles, on peut établir des règles pour passer automatiquement des documents sources aux documents cibles.
- L'intérêt est de repartir de l'original, non de la migration antérieure.

Pierre de rosette numérique

- Les connaissances stockées sur des supports restés lisibles permettent d'accéder aux contenus sur des médias obsolètes.
  - Etape de préservation de connaissance :
    - l'objectif est de disposer des connaissances / spécifications des contenus à conserver, sur plusieurs aspects: géométrie du stockage des données, méthodes de stockages, schémas d'encodage, modèle mémoire des fichiers, format des métadonnées ;
  - Extraction des données :
    - à l'aide des connaissances précédentes, récupérer les données issues des contenus conservés.
  - Reconstruction du contenu :
    - la connaissance préservée permet de reconstruire le document dans sa forme originale en utilisant des méthodes éventuellement totalement différentes des méthodes originales.

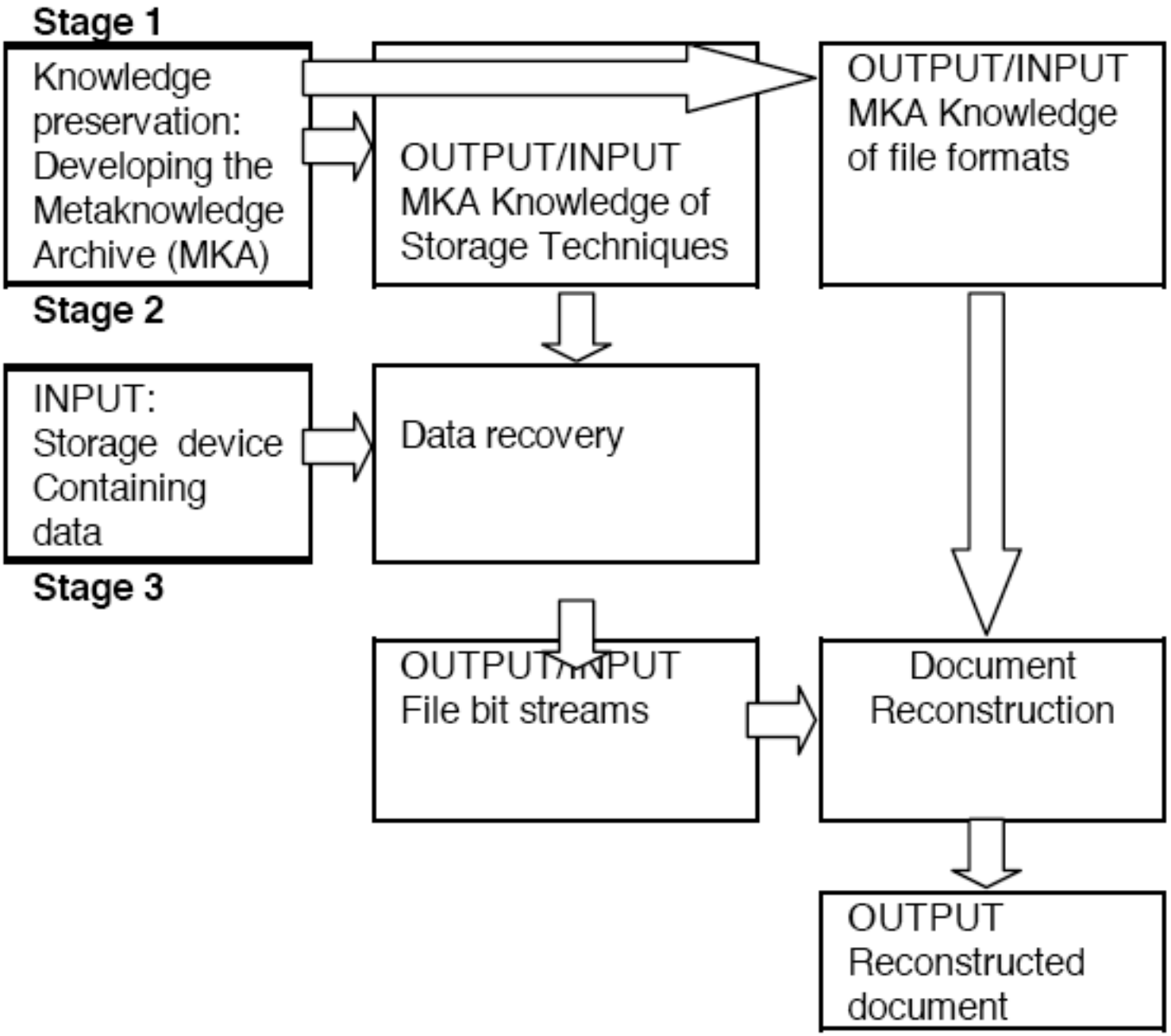


Figure 1. Digital Rosetta Stone Model

Statut de la description

- Plusieurs statuts:
  - Description de l'objet qu'il faut alors recréer de manière arbitraire au niveau technique, pour retrouver la description.
    - On ne sait pas comment il faudra faire, mais on sait ce que cela doit donner.
  - Description / prescription de ce qu'il faut faire pour recréer l'objet.
    - On ne sait pas ce que cela donnera au final, mais on a les instructions pour le faire.

Pertinence de la description

- La description doit capter « l'essence » du contenu ou de l'objet:
  - Est-il possible de cerner l'essence d'une œuvre ?
    - Donnée : oui, par principe. Mais la description est alors équivalente à l'objet.
    - Contenu : non, l'essence varie selon les époques et les contextes.

Description, interprétation

- Pour créer la description:
  - Il faut interpréter le contenu / l'objet à préserver et lui imposer une grille de lecture ;
- Pour exploiter la description:
  - La description est elle-même un contenu qu'il faut interpréter pour la mettre en œuvre.

Description : conclusion

- La description est une ré-invention :
  - On ne peut garantir l'identité à l'origine ;
  - Elle repose une compréhension et un culture du contenu permettant la création et l'exploitation des descriptions.
- Modèle qui tend à s'imposer :
  - Complexité technologique avec l'obsolescence de plus en plus rapide indique qu'il faut s'intéresser aux fonctions, informations et expressions, et moins au comment de l'outil technique pour préserver un contenu / objet.

DISTRIBUTION

Distribution

- Principe :
  - Les contenus sont des expressions permettant une communication et sont de fait appelés à circuler.
  - Les réseaux de communication contiennent le contenu à préserver.
  - Tout contenu communiqué est d'emblée conservé par le réseau de communication, à condition qu'on s'en serve.
- Démarche :
  - Ne pas conserver physiquement les contenus, mais les mettre en circulation.

Réalité de l'approche

- Conservation:
  - Publier un contenu
- Exploitation :
  - Accéder au contenu via le réseau et les mécanismes points à points (logique internet)
- L'Ina conserverait autant en mettant en ligne, et en se faisant pirater, qu'en numérisant et archivant les contenus…

Pertinence

- Cette approche est confirmée par l'histoire des contenus :
  - Plus un manuscrit était copié, plus il avait de chance d'être retrouvé dans le futur.
  - La démultiplication est un gage pour la préservation.

Fragilité -1: la durée

- Les points du réseau de communication ne pensent pas à la préservation, mais seulement à l'usage.
- Cet usage n'est pas provoqué, mais spontané.
- L'approche est donc fondée sur la compétition de l'usage et risque de perdre des contenus importants mais peu populaires.
  - Darwinisme de la préservation : Lady Gaga contre Kant ?

Fragilité – 2 : l'identité

- La communication est toujours une altération:
  - Chaque recopie est une falsification partielle de l'œuvre.
  - Le numérique est un support mutable par essence : toute recopie conduit à une réédition.
- La distribution doit se confronter à une multiplicité d'exemplaires voisins mais différents, différents mais voisins.

Fragilité – 3: l'authenticité

- La circulation ne permet pas de faire la part entre les transformations innocentes de celles qui faussent le contenu.
  - Rééditions qui reconditionnent le contenu;
  - Malversations qui faussent le contenu.
- L'enjeu est de permettre de faire la part entre ces différents cas.

Centraliser l'identité

- La distribution doit donc être pensée avec des outils permettant d'établir la référence et l'identité des contenus circulant dans le réseau.
- Fonction possible des institutions de mémoire:
  - Pouvoir identifier et certifier les contenus;
  - Interpréter les variantes et exemplaires;
  - Remettre en circulation les contenus importants.

Approche coopérative

- La distribution n'est pas une approche exclusive, mais doit être pensée avec les approches centrées sur la préservation centralisée des contenus.
- La distribution ne dit pas que la centralisation est inutile, mais indique plutôt que la distribution est une composante de la stratégie de préservation des institutions centralisées de la mémoire:
  - L'archive centralisée centralise la référence et l'identification;
  - L'archive préserve un exemplaire témoin mais démultiplie la préservation par la mise en circulation.

Des tentatives en cours

- Cloud preservation
- LOCKSS

Le cloud

Un service, pas une technologie en soi

Derrière le cloud…

HDD

SSD

LTO

Des offres à foison…

Avantages
Sécurité du stockage
Pas de coût de migration direct
Rapidité de mise en œuvre (15 mns)

	Access mode	Storage 1 TB / 1 month	Access 10 GB / month	Annual cost	Adapted to
Multi Regional	+++ Très rapide Depuis n'importe où	+++ 26 $	-- 1.20 $	327 $	Livraison "client"
Regional	++ Très rapide Limité au continent	++ 20 to 23 $	-- 1.20 $	290 $	Livraison locale de travail
Near line	- Accès mensuel	- 10 $	+ 2.20 $	147 $	Usage de la plupart des structures d'archives
Cold Line	-- Accès annuel	-- 7 $	++ 6.20 $	158 $	Backup, sécurité

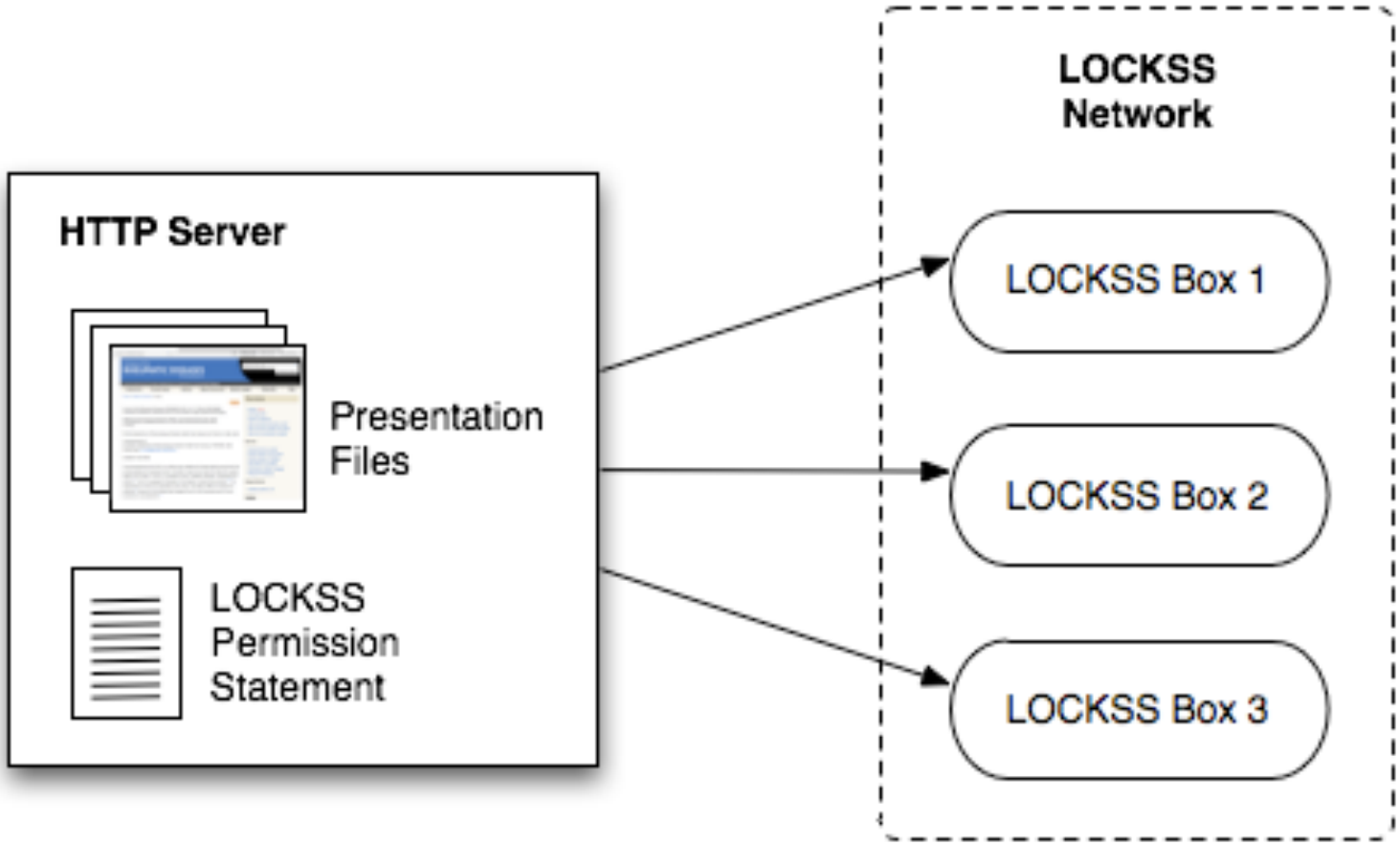
LOCKSS

- Projet de déjà 15 ans porté par Stanford.
- « open source, peer-to-peer, decentralized digital preservation infrastructure »
- « all formats and genres of web-published content »
- Conforme à OAIS
- Audit continuel des octets pour réparation et migration par comparaison des multiples archives (LOCKSS Boxes) entre elles

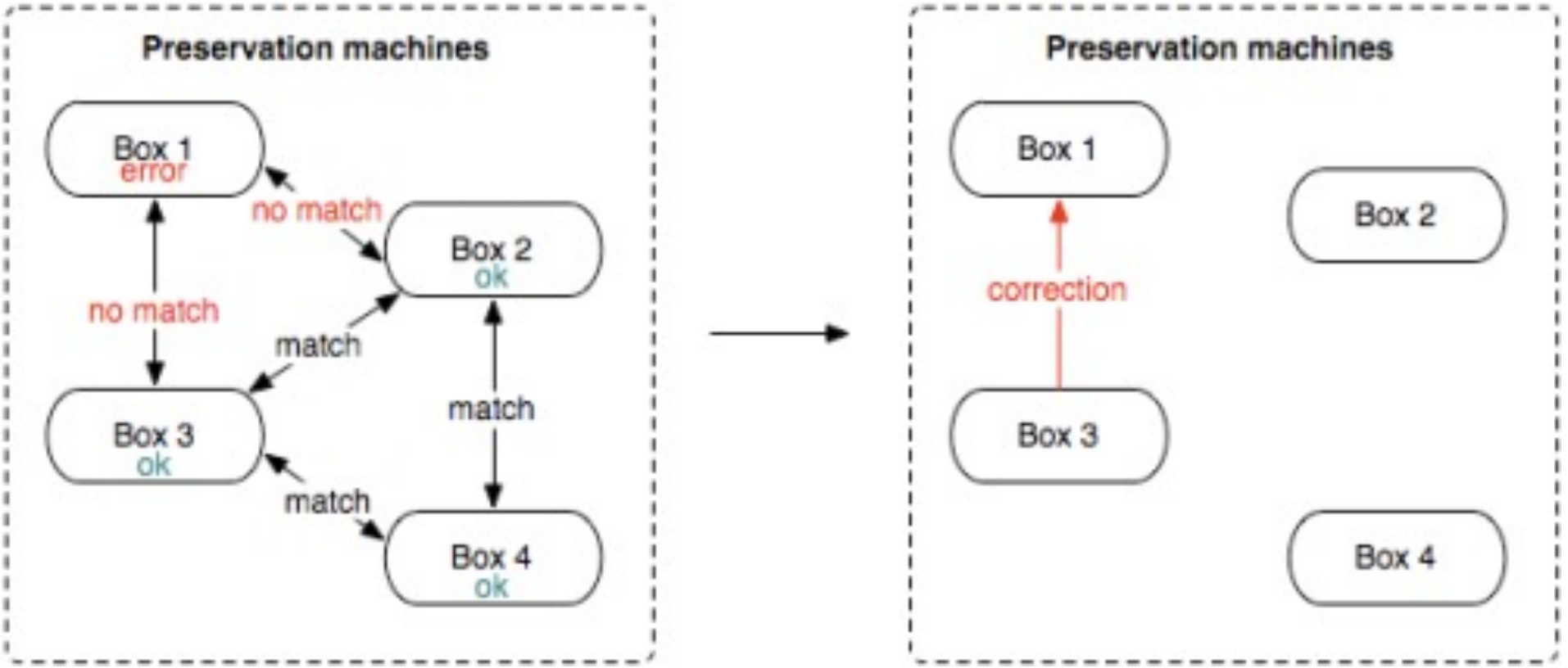
Collecte

- Chaque bibliothèque maintient une « LOCKSS box » qui va archiver une série d'adresse d'éditeurs
- Chaque site d'éditeurs autorise sa captation par les robots LOCKSS
- Chaque boîte constituée sont ensuite comparées et harmonisées pour définir la « bonne » version à partir de celles qui sont été captées.

Collecter l'information



Maintenir à partir d'un réseau de LOCKSS boxes



Accès à l'information

- Conservation dans le format d'origine
  - Pas de copies migrées à conserver
  - Une seule version de référence
- Migration à la volée lors de l'accès au contenu selon les contraintes des browsers (Migration on Access)
  - On part toujours de la « bonne version » de référence
  - On bénéficie de la meilleure technologie du moment
  - Les convertisseurs sont conservés et associés aux contenus originaux et leurs formats.

LOCKSS aujourd'hui

GLN

- réseau global LOCKSS:
  - GlobalLOCKSS Network – GLN
- préservation des publications électroniques scientifiques(e-journaux, e-books) qui intéressent suffisamment de participants
  - les «collections générales » des bibliothèques

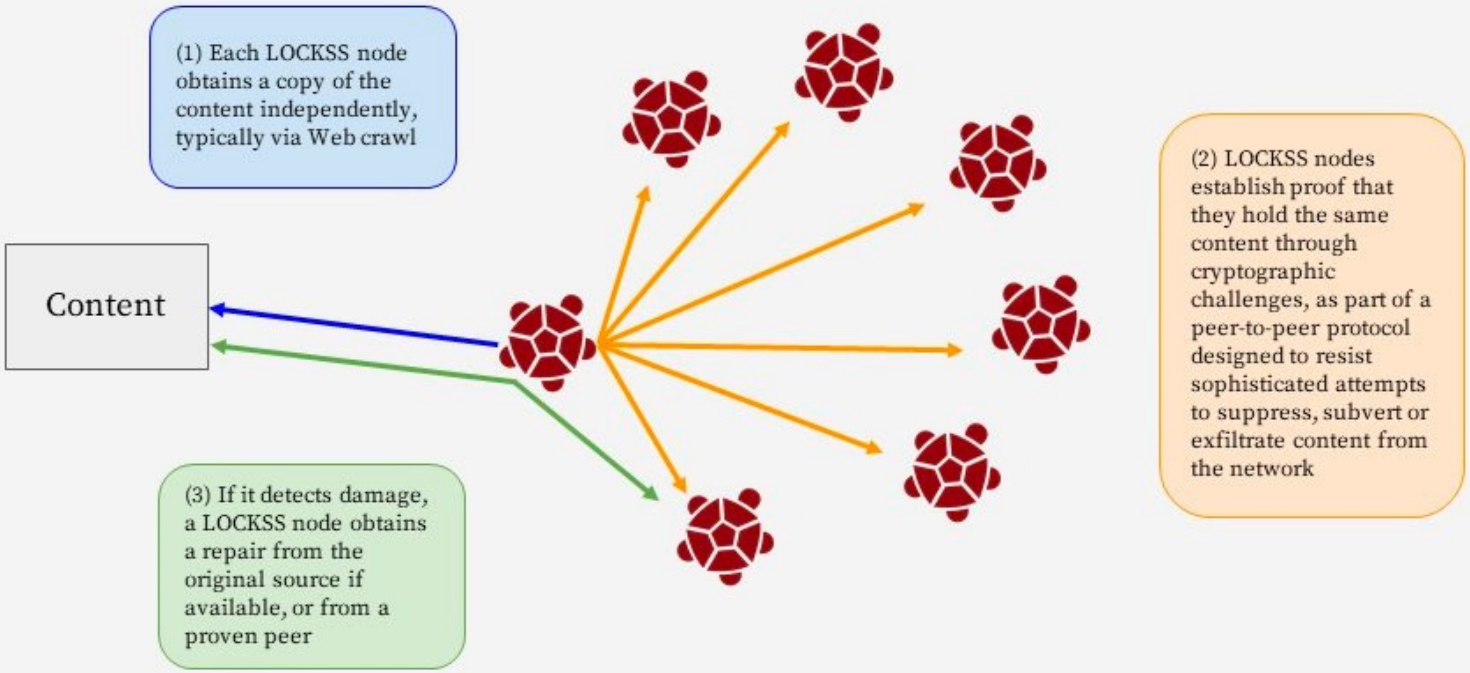
PLN

- Private LOCKSS Network – PLN,
- contenus numérisés ou natifs qui appartiennent à une communauté spécifique(collections audio, images, databases, documents…)
- contenus éditeurs qu'une communauté souhaite préserver indépendamment du Global LOCKSS Network
  - pour des raisons de contrôle, de propriété des données ou de spécialisation

GLN

Stanford | LIBRARIES

Visual Overview of a LOCKSS Network



LOCKSS CLOCKSS

2

Deux niveaux: Préservation et l'accès perpétuel

Dark Archive

- L'archive est destinée uniquement à l'archivage pérenne et soumise à des règles d'accès strictes.
- Accès uniquement si les contenus ne sont plus disponibles ailleurs

Light Archive

- combine les fonctions de préservation à long terme et de communication
- propose un accès plus ouvert

Trois solutions pour les e-journaux

- LOCKSS(GLN)
  - Logiciel OS (Stanford), réseau distribuéde préservation, ouvert;
  - 600 éditeurs, 150 bibliothèques, 11 000 titre
  - Souscription annuelle (3 à 10 000 € + 1000€ matériel + 2400 gestion)
- CLOCKSS (PLN)
  - 11 éditeurs et 12 bibliothèques, système distribué, coûts réduits
  - Logiciel OS (Stanford) + PLN
  - Pas d'accès post-abonnement ni en cas de transfert de titre
  - 201 éditeurs, 271 bibliothèques, 14 000 titres, 400 – 1800€
- PORTICO
  - Archive centralisée + Sites miroirs,logiciel propriétaire
  - 334 éditeurs, 996 bibliothèques, 22 000 titres.
  - 3 000 à 10 000 €, éditeurs : 250 à 80 000 €

Etre ou ne pas être Lockss

SOLUTIONS LIÉES A LOCKSS

- **UK LOCKSS Alliance(GLN)**
Créée en 2008, c'est réseau LOCKSS de bibliothèques universitaires anglaises participant au GLN. 15 institutions y participent.

- **LUKII (Allemagne)(PLN)**
Le projet allemand LUKII (2010-2012) visait à combiner les forces de LOCKSS et de KoLiBRI, logiciel Open Source développé entre 2004 et 2007 dans le cadre du projet KOPAL. Le projet était constitué de 9 LOCKSSBOX sous la forme d'un PLN.

- **Près de 15 PLN actuellement**

France : Quelques éditeurs participent mais aucune bibliothèque

SOLUTIONS HORS LOCKSS

- **Dépot légal électronique**
A pour but de garantir la préservation à très long terme des publications électroniques.
Mais ne prend en compte seulement les publications nationales. De plus, les bibliothèques nationales et universitaire n'ont pas les même priorités en terme d'accès au documents.

-**E-Depot(Suisse, cloturé en 2011)**
A partir de 2011, une LOCKSBOX + 6 bibliothèques participent à PORTICO

- **SPAR (France,Bnf)**
En 2013, la Bnf ouvre son système d'archivage numérique SPAR (Système de Préservation et d'Archivage Réparti), lancé en 2010, à d'autres organisations, qui peuvent ainsi bénéficier de l'expertise et des infrastructures de la BnF, via le service de « tiers archivage ».

Conclusion

- L'objet à préserver n'est jamais auto-explicatif
  - La description est donc une composante nécessaire de la préservation, et parfois la composante essentielle et la seule faisable.
- L'objet à préserver doit résister aux aléas des institutions et des erreurs / désastres politiques
  - La circulation des contenus est une composante nécessaire et complémentaire de leur conservation centralisée.