Huma-Num
la TGIR des humanités numériques

Archivage

Michel Jacobson
Huma-Num

ANF « Gestion de projet sur les sources numériques en SHS »
Aussois - 6-8-octobre-2020

cnrs Aix*Marseille université CAMPUS CONDORCET Paris–Aubervilliers

Archivage

- Plan
  - Définitions
    - Archives
    - Information
    - Cycle de vie
  - Quelques particularités du numérique
  - Les grandes fonctions d'un Système d'archivage électronique

A(a)rchive(s)

- « *Les **archives** sont l'ensemble des documents, quels que soient leur date, leur forme et leur support matériel, produits ou reçus par toute personne physique ou morale et par tout service ou organisme public ou privé dans l'exercice de leur activité.* »
  - *Code du patrimoine, article L. 211-1.*

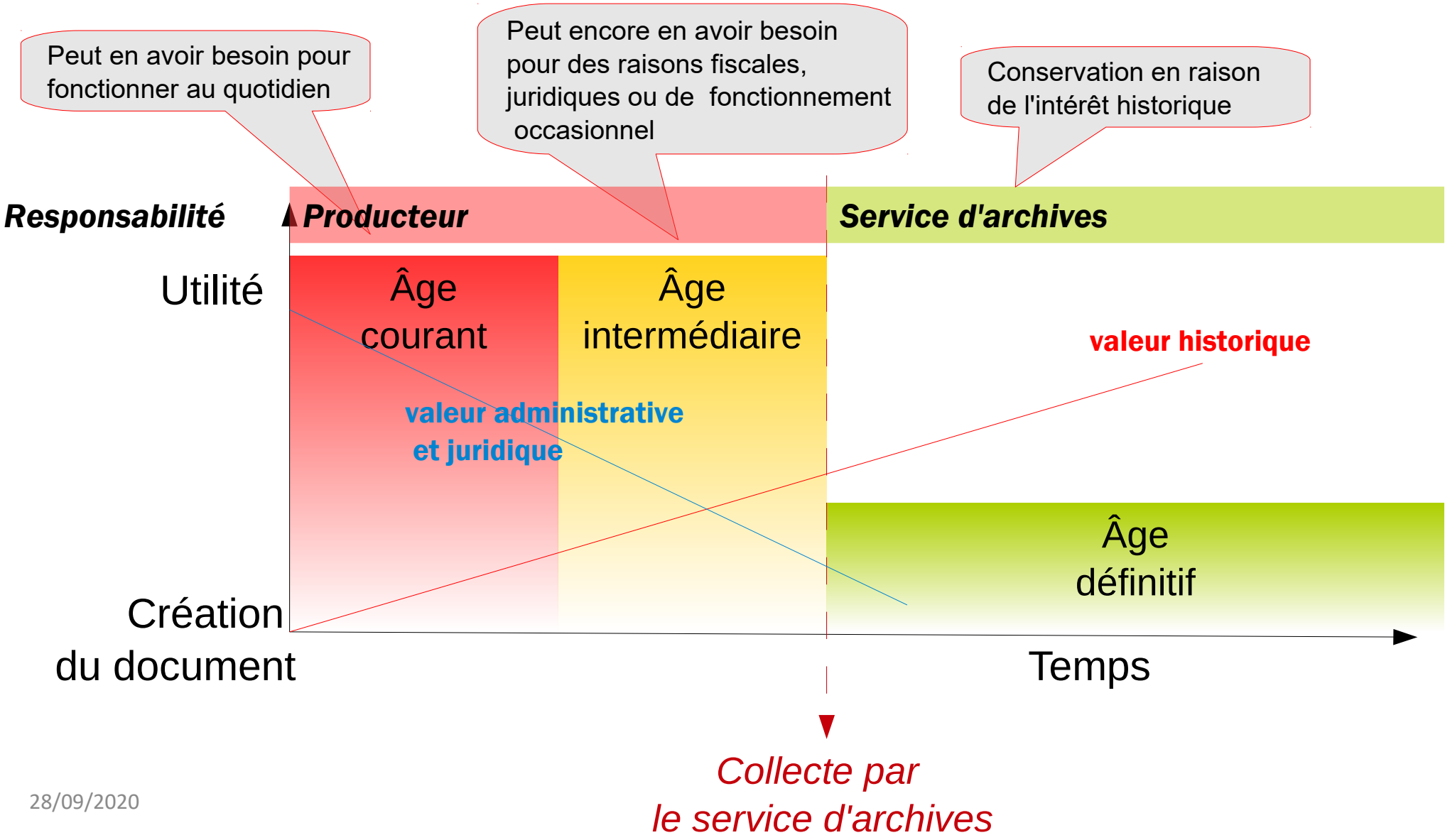
- « *Une **Archive** est une organisation dont la vocation est de préserver l'information pour permettre à une communauté d'utilisateurs cible d'y accéder et de l'utiliser.* »
  - *Norme ISO 14721:2003 (révisée 2012) « Open archival information system »*

L'information

- La norme OAIS n'utilise que rarement le terme document, au bénéfice du terme information. L'OAIS distingue :
  - l'information: connaissance que l'on peut échanger ;
  - la donnée: représentation formalisée de la connaissance.
- Pour conserver une information, on lui donne une forme (la donnée). Cette forme peut être liée à des technologies qui sont éphémères. On a donc besoin de conserver aussi des informations sur cette forme (les métadonnées)

Organisation internationale de normalisation

Le cycle de vie de l'archive



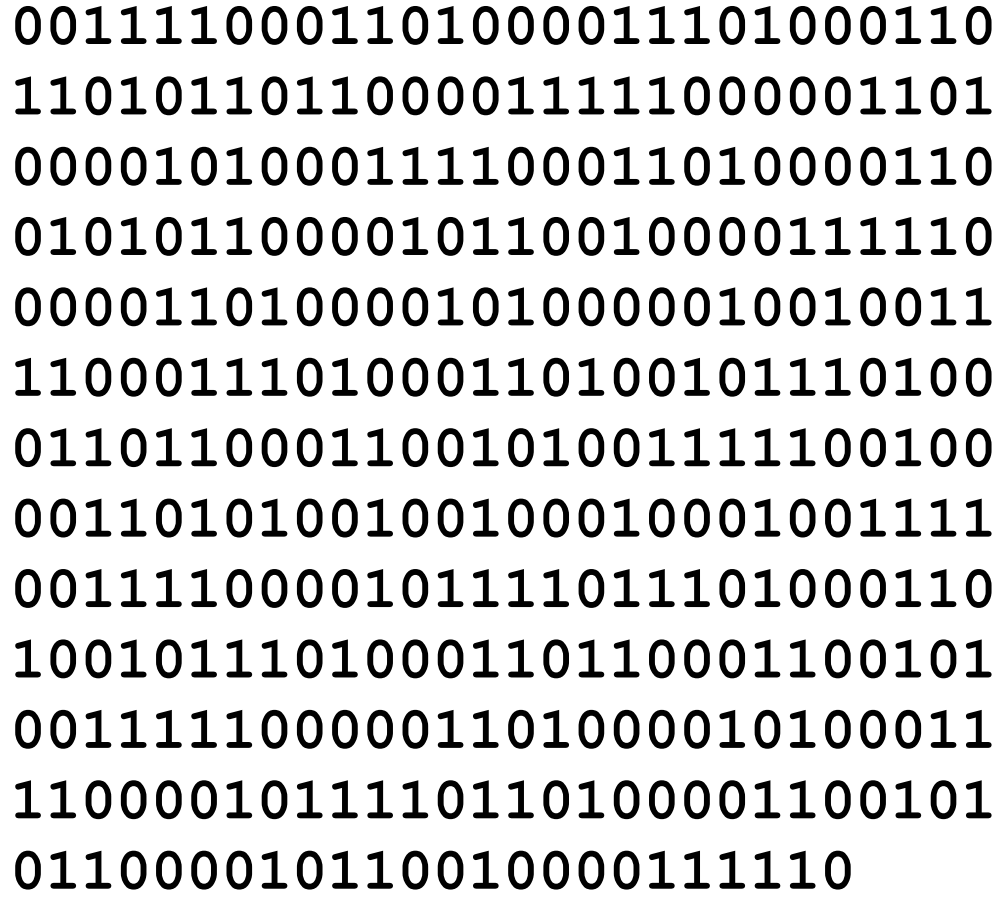
Le sort final

- Gestion du cycle de vie de l'information
  - La durée d'utilité couvre les deux premiers âges.
  - A l'issue de cette période le producteur applique un sort final aux documents.
    - Cas général : si le producteur n'a plus besoin de l'information, il peut la détruire
      - Pour les archives publiques, il doit obtenir un visa d'élimination délivré par l'autorité de contrôle compétente.
    - Cas particulier : en raison d'un intérêt historique, scientifique, statistique, le code du patrimoine permet le changement de finalité et le transfert de responsabilité à un service public d'archives.

Quelques particularités de l'information numérique

La représentation numérique

- Suite organisée d'unités binaires
- Touche tous les domaines
- Reproduction à l'identique
- Reproduction à l'infini
- Stockage à faible coût
- Facile à transmettre
- Facile à traiter
- Mais difficile à conserver...



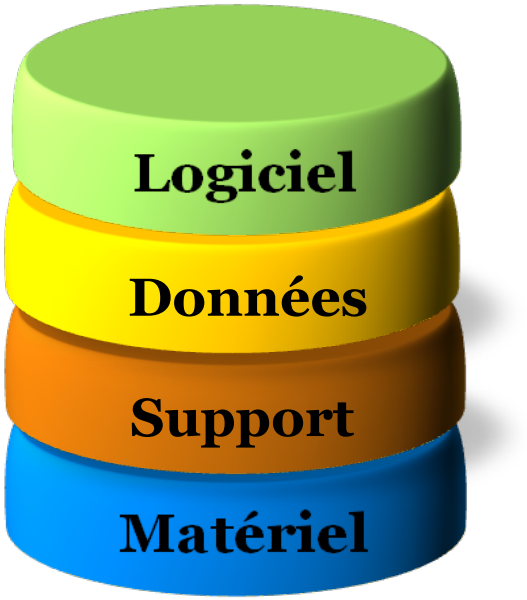
Vulnérabilité de l'information numérique

- Mais...

  l'information sous forme numérique est difficile à conserver car :

  – Elle est codée

  – Elle dépend d'une pile de couches interdépendantes de technologies diverses



Vulnérabilité du numérique

- Vieillissement (supports, matériels)
- Obsolescence (lecteur, connectique, logiciel (driver, système d’exploitation, logiciel applicatif, format, codage
- Facteurs de marché (arrêt de la fabrication, de la maintenance, disparition des compétences et des pièces détachées)
  - Conséquences : perte d'information brutale, massive, qui peut passer complètement inaperçue.

Les grandes fonctions d'un système d'archivage électronique (SAE)

---

intégrité, lisibilité, traçabilité

Intégrité

- Objectif : garantir que le contenu informationnel n'a subi aucun altération
- Technique : calcul/comparaison d’empreintes numériques
- Quand est-elle mise en œuvre ?
  - En cas d'opérations de réécritures (copies pour communication, migrations de supports, etc.)
  - Pour surveiller le vieillissements des supports
  - Pour repérer les doublons

Lutte contre la perte d’intégrité

- Redonder l’information
  - prévue par le support (CD) ou par le logiciel ou le matériel (RAID) [vieillissement, panne]
  - plusieurs lieux de stockage distants de plusieurs centaines de mètre (ou km) [incendie, inondation, etc.]
  - plusieurs types de supports [facteurs de marché, lutte anti virale…]
  - un système de sauvegarde [continuité de service]

Toutes ces mesures permettent en cas de corruption avérée de restaurer la donnée à partir d’une copie intègre

Lisibilité

- Objectif : que l'interprétation de l'information reste possible et correcte
- Maintient de la lisibilité à travers :
  - La migration de formats
    - Pour lutter contre l'obsolescence
      - Conversion des encodages de textes EBCDIC → ASCII → UTF-8
    - Pour lutter contre la dépendance vis à vis du marché ou pour maîtriser toutes les métadonnées nécessaire à l'interprétation
      - Conversion format propriétaire → format libre (DOC → ODT)
    - Pour lutter contre la dépendance vis à vis de l'environnement système
      - Conversion vers des formats autoporteurs (PDF/A)
    - Pour permettre de nouvelles fonctionnalités
      - PDF/A-1 → PDF/A-2 (+ gestion de la transparence)
      - WAVE → BWF (encapsulation de métadonnées)
- Une autre solution pour le maintient de la lisibilité : l'émulation

Tracabilité

- Établir une chaîne de confiance
  - Tous les événements sont datés et enregistrés dans un journal
    - C'est lui qui fera foi pour le test de comparaison d'empreinte.
    - Les migrations de formats qui entraînent une rupture d'intégrité peuvent ainsi être traçées jusqu'à remonter à l'entrée d'une donnée dans le système.
  - Les journaux sont archivés de la même manière que les autres documents
    - C'est-à-dire qu'on calcule aussi leurs empreintes et que l'on écrit ces empreintes dans les journaux.

Fonctionnalités des SAE

- Un SAE comporte aussi d'autre fonctions
  - Fonctions d'identification, d'authentification, de conversion, de restriction d'accès, de supervision, de classement, de gestion du cycle de vie, de contrôle, de recherche, etc.

Comment bien gérer ses données ?

- Identifier et caractériser tous les types de données manipulés dans le projet
- Définir le cycle de vie de chaque type de données
- Déterminer les moyens à mettre en œuvre pour la gestion de chaque type de données à tous les âges de l'archive
  - Trouver les bonnes solutions en termes de locaux, matériels, logiciels, formats, codages, services...

Cf. DMP

Conclusion

- Archiver c'est mettre en place une gestion rationalisée des informations
  - Ça commence à la naissance de l'information
  - Ça continue tout au long de son cycle de vie jusqu'à son élimination ou parfois sans limite de temps
  - C'est lutter contre les risques de perte d'intégrité (vieillissement, falsification), de lisibilité (obsolescence), d'accessibilité (manque de métadonnées)