ESI/HEG/ENC – École d'été « Patrimoine et numérique » - Rabat, 9-15 juin 2022

Enjeux de numérisation et de préservation : institutionnels, historiques, mémoriels et techniques

Rabat, 12 juin 2022

École nationale des chartes | PSL

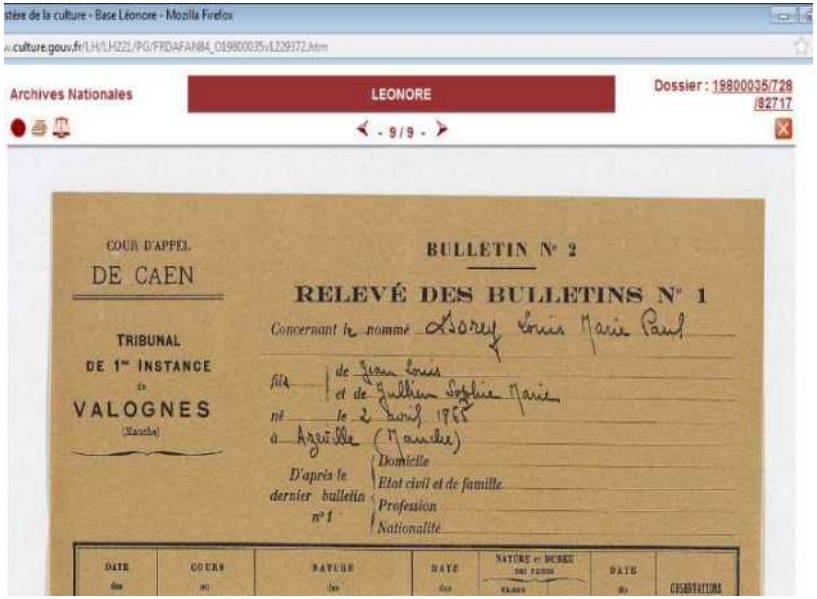
Qui suis-je ?

Édouard Vasseur

- Diplômé de l'École nationale des chartes puis de l'Institut national du patrimoine
- Docteur en histoire
- Pendant 17 ans, conservateur du patrimoine, spécialité archives :
  - Responsable de la gestion des fonds contemporains aux Archives nationales
  - Responsable de la mission archives du ministère de la Culture
  - Chargé de mission archivage électronique au ministère de la Défense
  - Responsable fonctionnel du programme d'archivage électronique VITAM
- Depuis 2019, enseignant-chercheur à l'École nationale des chartes, titulaire de la chaire d'histoire des institutions, de diplomatique et d'archivistique contemporaine

Introduction

Deux types de patrimoines



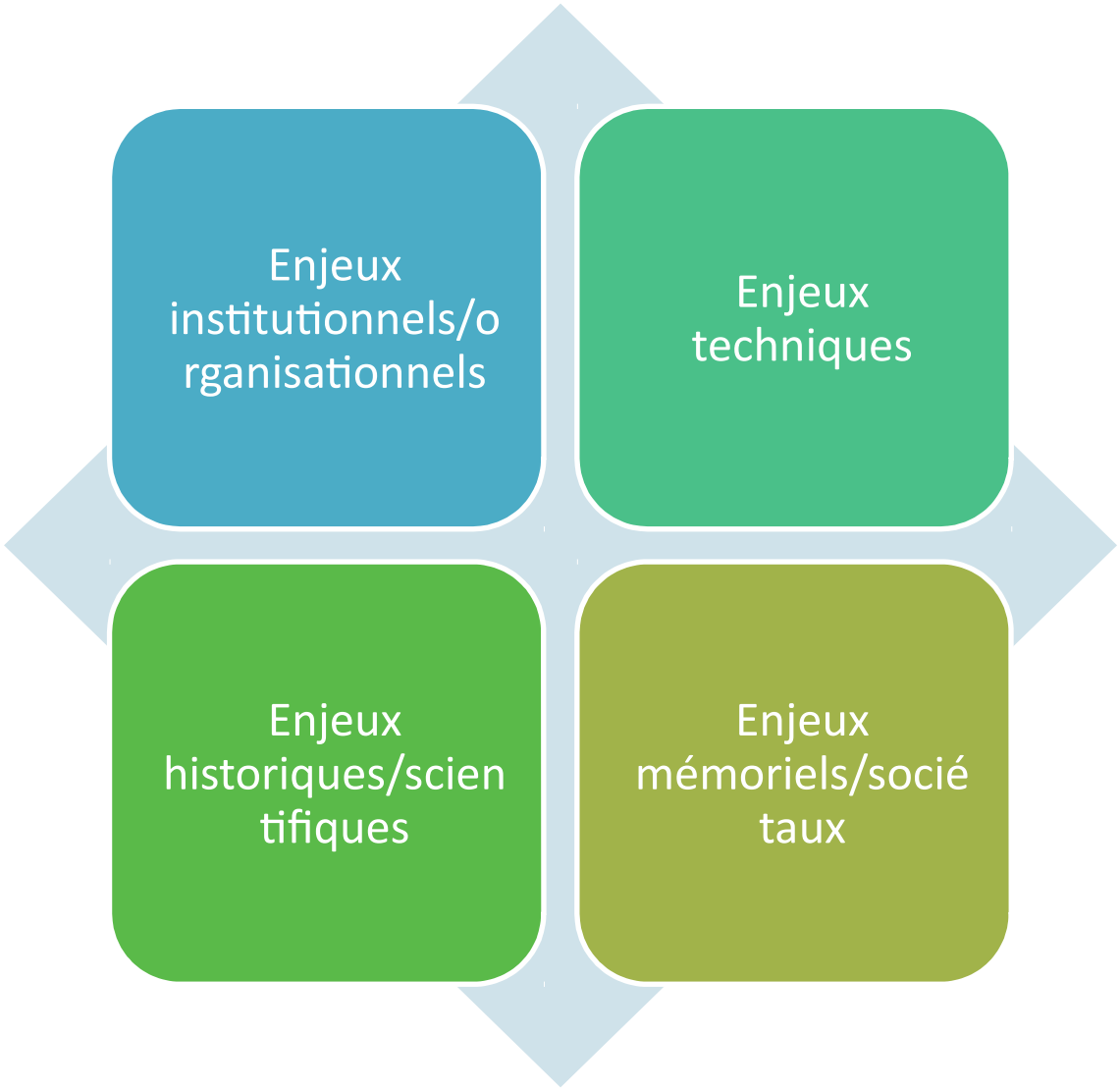
Le patrimoine numérisé :

- Existe déjà sur un autre support physique, qui continue à exister en parallèle de la version sur support numérique
- Est le résultat d'une action délibérée et organisée de création

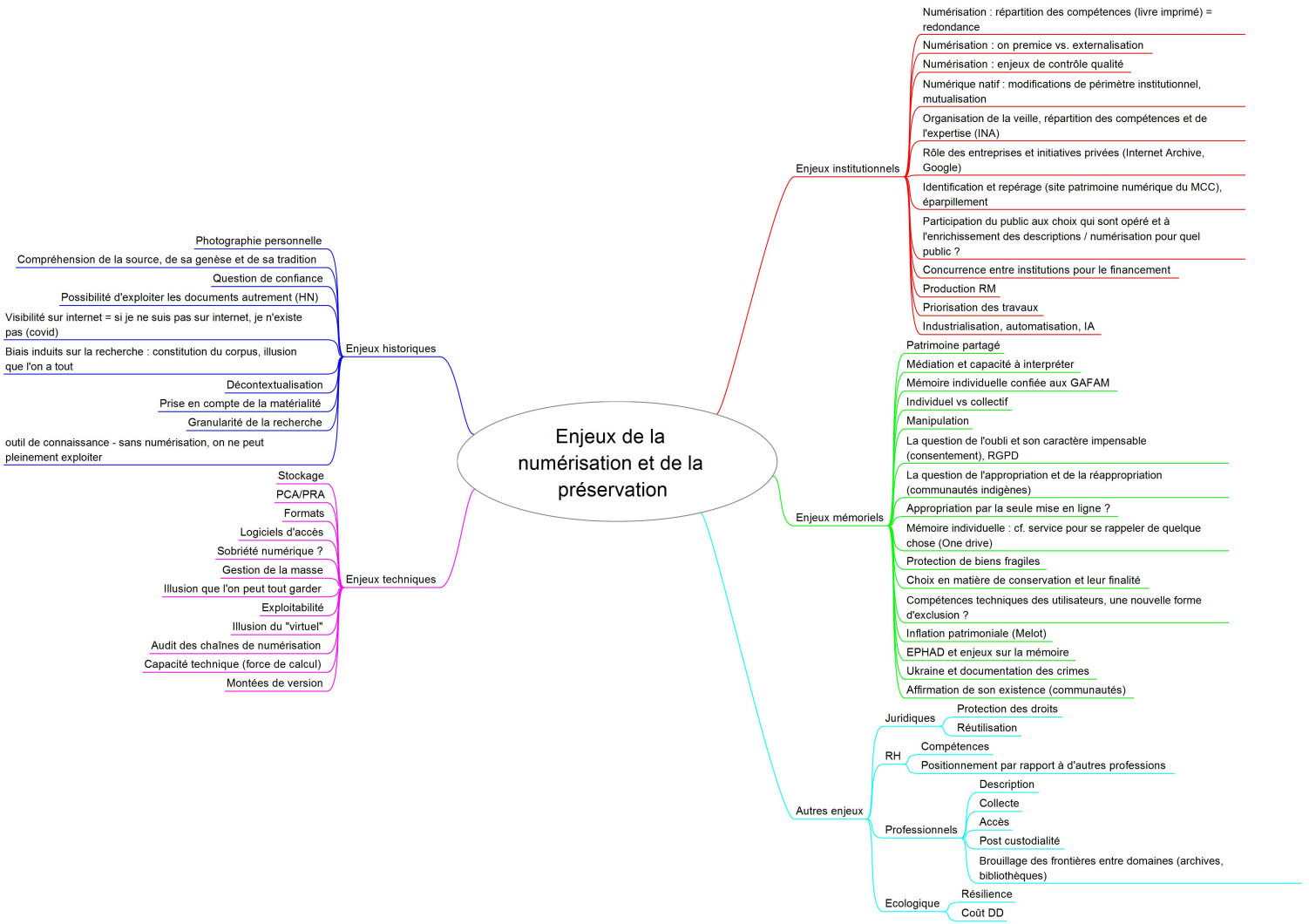
Le patrimoine nativement numérique

- N'existe que sur ce support
- Est créé a priori sans l'intervention des acteurs du patrimoine

Quatre enjeux à examiner



Comment procéder ?



Plan de l'intervention

- Le patrimoine numérisé
- Le patrimoine nativement numérique
- Conclusion : quels sont les points communs et les différences entre les deux types de patrimoine ?
- Discussion

Patrimoine numérisé

Enjeux institutionnels/organisationnels - 1

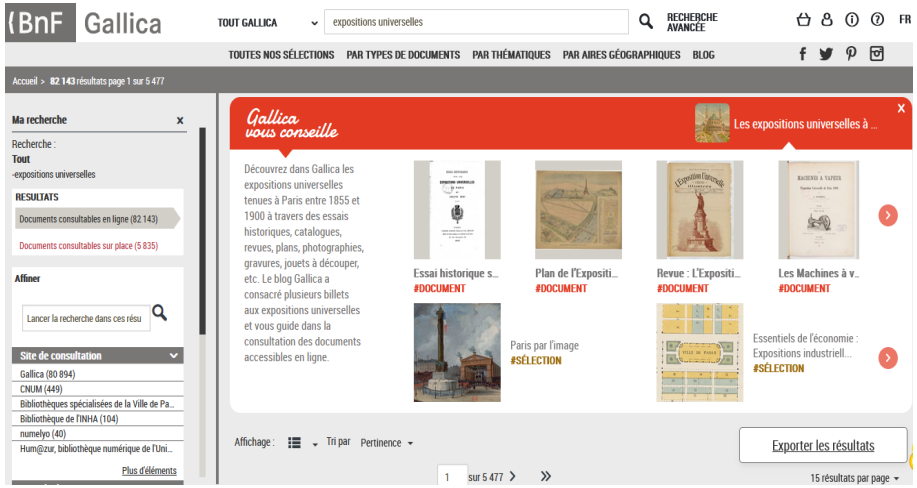
Plusieurs sujets:

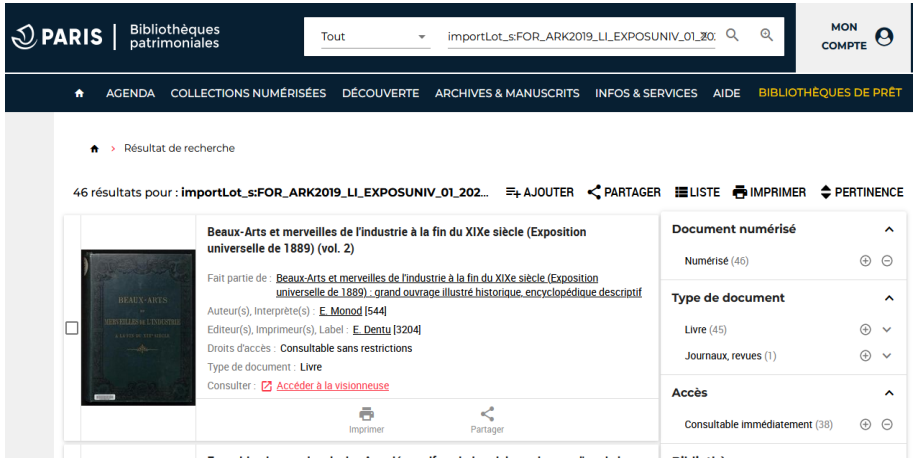
- - **méthodologie de réalisation** : réalisation en interne ou recours à des prestations
- - **définition des standards de numérisation**, dans tous leurs aspects (définition de la numérisation, format des fichiers, nommage des fichiers, métadonnées à générer)
- - quelle que soit la solution retenue, **organisation des chaînes de numérisation** : priorisation, identification des corpus à numériser, préparation des corpus à numériser (matérielle comme intellectuelle), transfert à l'entité chargée de procéder à la numérisation, retour des corpus à numériser et réception des versions numérisées, contrôle des versions numérisées et des métadonnées générées, intégration des versions numérisées dans le système d'archivage (d'un point de vue global)

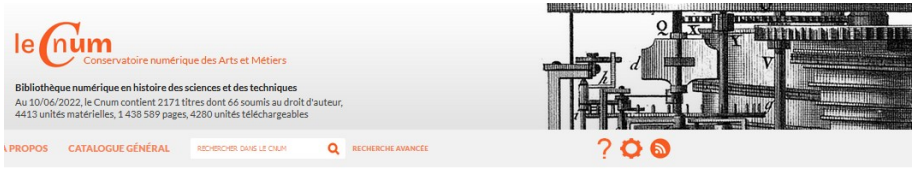
Et une grande question : le financement !

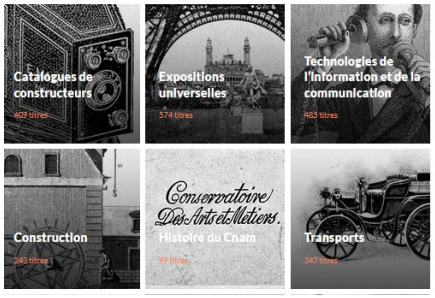
Enjeux institutionnels/organisationnels - 2

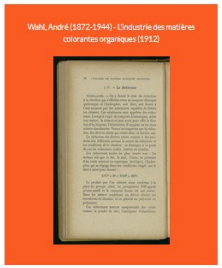
Comment éviter les redondances ? L'exemple du corpus des Expositions universelles











Enjeux institutionnels/organisationnels - 3

Quelle place pour le secteur privé ? L'exemple des relations avcec Google

Enjeux institutionnels/organisationnels - 4

Quelle place pour les utilisateurs (« les clients ») ?

ARCHIVES DÉPARTEMENTALES DES YVELINES

Yvelines Le Département | NOUS CONNAÎTRE | RECHERCHER | PARTICIPER | EDUC'ARCHIVES

Accueil > Participer > A la recherche de la presse perdue

- Jeu Gueule d'Ange
- Adoptez un poilu !
- Le WIKI de la Grande Guerre
- Testaments de Poilus | Saison 2
- C'est quoi ce chantier ?
- A la recherche de la presse perdue
- Jeu Réseaux

A la recherche de la presse perdue

Aidez les Archives à compléter les colle

Enjeux historiques/scientifiques - 1

Le chercheur en SHS dans les années 1990

Corpus disponibles :

- Des contenus enregistrés sur des supports physiques : parchemin, papier, disques, pellicules, diapositives
- Des supports de substitution sous forme de microfiches et microfilms
- Des reproductions possibles sous forme de photocopies
- Des systèmes de prêts à distance

Outils pour identifier les corpus:

- Catalogues et inventaires papier
- Quelques bases de données accessibles sur site ou consultables sur Minitel
- Des bibliographies au format papier (Bibliographie d'histoire de France)

Le chercheur en SHS dans les années 2020

Corpus disponibles :

- Les mêmes contenus que précédemment + sources nativement numériques
- Des corpus numérisés et disponibles directement sur internet
- Des fonctionnalités d'export

Outils pour identifier les corpus :

- Catalogues de sources en ligne
- Des bouquets de périodiques accessibles en ligne gratuitement (Persée) sur abonnement (Cairn, etc.)
- Des moteurs de recherche
- Des portails : ex. Isidore
- Des bibliographies accessibles directement en ligne (BHA)

Enjeux historiques/scientifiques - 2

Comprendre la constitution du corpus. L'exemple de la Burney Newspaper Collection de la British Library

Présentation de la Burney Newspaper Collection :

- Rassemblement par Charles Burney de périodiques et brochures disponibles dans des pubs londoniens
- Reliure en 700 volumes chronologiques

Les conditions de constitution du corpus numérique :

- Une collection complétée après acquisition par la BL
- Microfilmage réalisé par titre et non par volume
- Une numérisation à partir des microfilms mis en salle de lecture en non à partir des masters
- Une reconnaissance de caractères de mauvaise qualité

Pour en savoir plus : Katie Lanning (2020) Scanner darkly: unpopularization in the Burney Newspaper Collection, Archives and Records, 41:3, 215-235, DOI: 10.1080/23257962.2020.1810004

Enjeux historiques/scientifiques - 3

Comprendre la genèse du corpus - L'impact de la transposition de la collection d'interviews de survivants des camps de concentration rassemblée par le psychologue David Boder en 1946 en un site internet interactif en 2009 (Université du Luxembourg)

Enjeux historiques/scientifiques - 4

La numérisation permet-elle de comprendre et de restituer le contexte de création du corpus d'origine ?

Enjeux historiques/scientifiques - 5

La numérisation permet-elle d'appréhender la matérialité (voire la sensorialité) du patrimoine et peut-elle remplacer le contact direct avec lui ?

Léonard de Vinci, *Mona Lisa,* Paris, musée du Louvre

Véronèse, *Les noces de Cana,* Paris, musée du Louvre

Enjeux historiques/scientifiques - 6

Ce qui est numérisé est-il exploitable ?

Identification des corpus :

- En plus des propositions des institutions patrimoniales, masse des numérisations réalisées dans le cadre de projets de recherche

Exploitation de la masse :

- L'appareil photo numérique et le téléphone portable : les deux outils principaux du chercheur en archives
- Conséquence : des photographies qui s'accumulent sur tous les supports de stockage. Pour plus tard ...

Exploitation du contenu :

- Enjeux des nouvelles opportunités d'analyse offertes par les humanités numériques et l'intelligence artificielle (fouille de données, analyse sémantique)
- Des procédés de numérisation qui ont évolué dans le temps (passage du mode image au mode OCR)

Enjeux mémoriels/sociétaux - 1

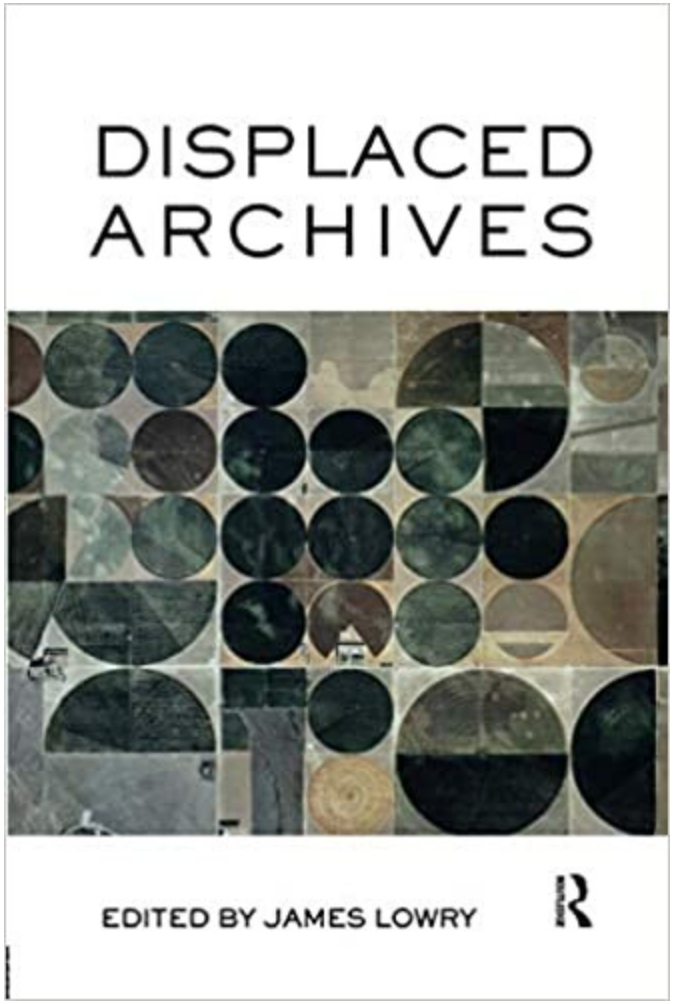
Le patrimoine numérisé, une chance pour le patrimoine partagé et le patrimoine en danger ?

Le patrimoine, aux aléas de l'histoire :

- Saisies de guerre
- Dominations impériales et coloniales
- Marché des biens culturels
- Disparition de locuteurs
- Conséquences des événements climatiques

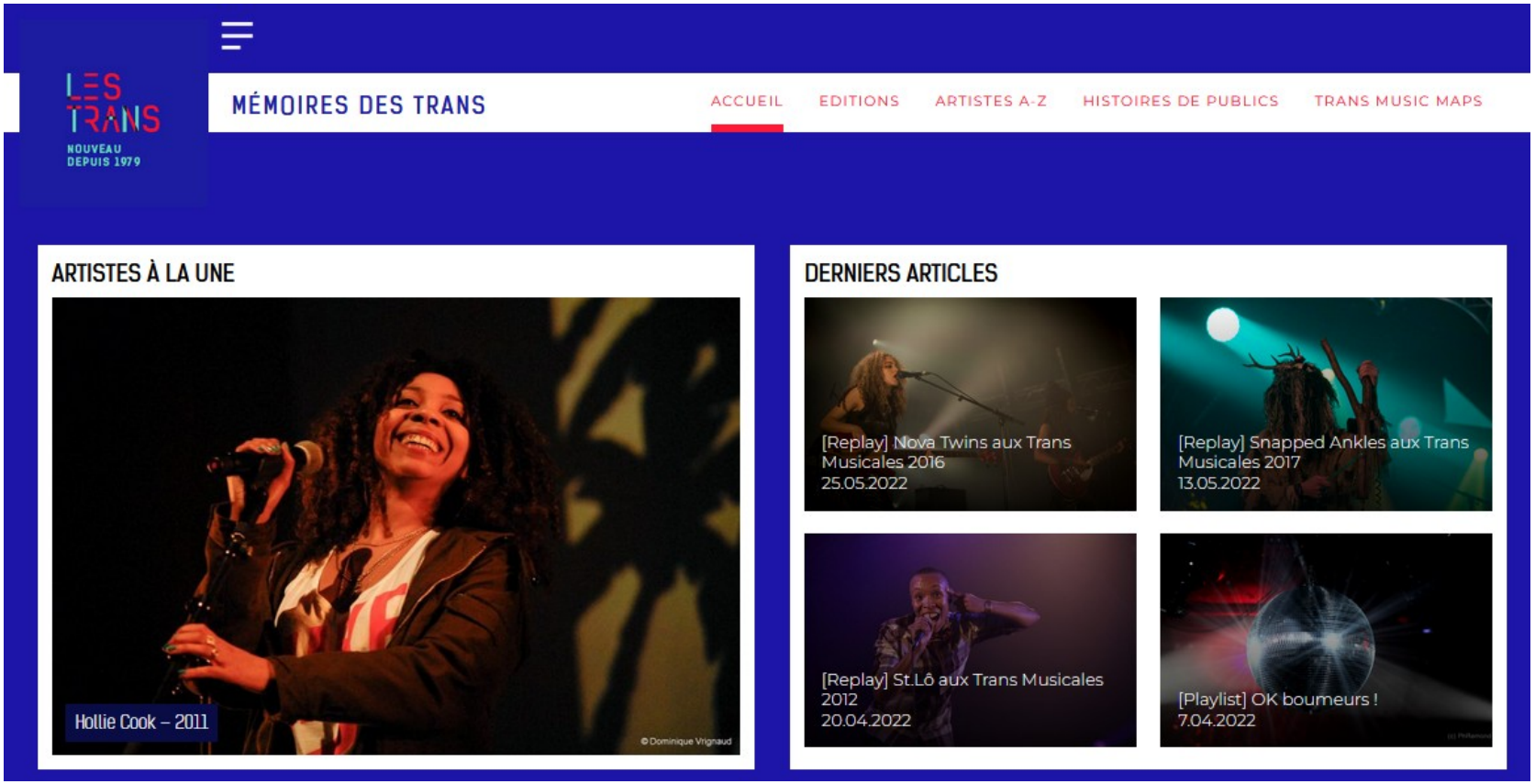
La numérisation, un moyen de pallier ces aléas ? :

- L'expérience du microfilmage
- Grâce à l'internet et au web, des moyens d'accès renouvelés et facilités
- L'enjeu de l'appropriation et de la réappropriation par les populations concernées



Enjeux mémoriels/sociétaux - 2

Le patrimoine numérisé, un enjeu d'existence pour les communautés ?



Enjeux mémoriels/sociétaux - 3

Le patrimoine numérisé, un outil pour favoriser le lien social ?

Sur les chemins de la mémoire

Les Archives départementales de l'Ardèche ouvrent leurs fonds et leurs portes aux animateurs et aux résidents d'établissements pour personnes âgées et les invitent à partager la mémoire et le patrimoine du département.

Les **activités** proposées ont pour objectifs de partager un moment de convivialité en évoquant ses souvenirs, de favoriser les échanges et de maintenir du lien social. Au nombre de trois, elles peuvent également prendre la forme d'une rencontre intergénérationnelle : visite découverte des Archives, animations (sur place ou en établissement) et prêt de jeux.

Enjeux mémoriels/sociétaux - 4

Le patrimoine numérisé, un enjeu en terme d'évolution des professions patrimoniales ?

Une remise en cause des professions ?

- Quelles conséquences sur les priorités d'acquisition, de conservation et de traitement ?
- Quels besoins en terme de médiation ?
- Accepter la perte de maîtrise sur la réutilisation et les réutilisations/réappropriations inappropriées ?

Un nouveau risque d'exclusion ? :

- Incapacité à accéder à internet ou à utiliser les services offerts ?
- Invisibilisation sur le net vs. Surreprésentation de certaines communautés

Enjeux techniques

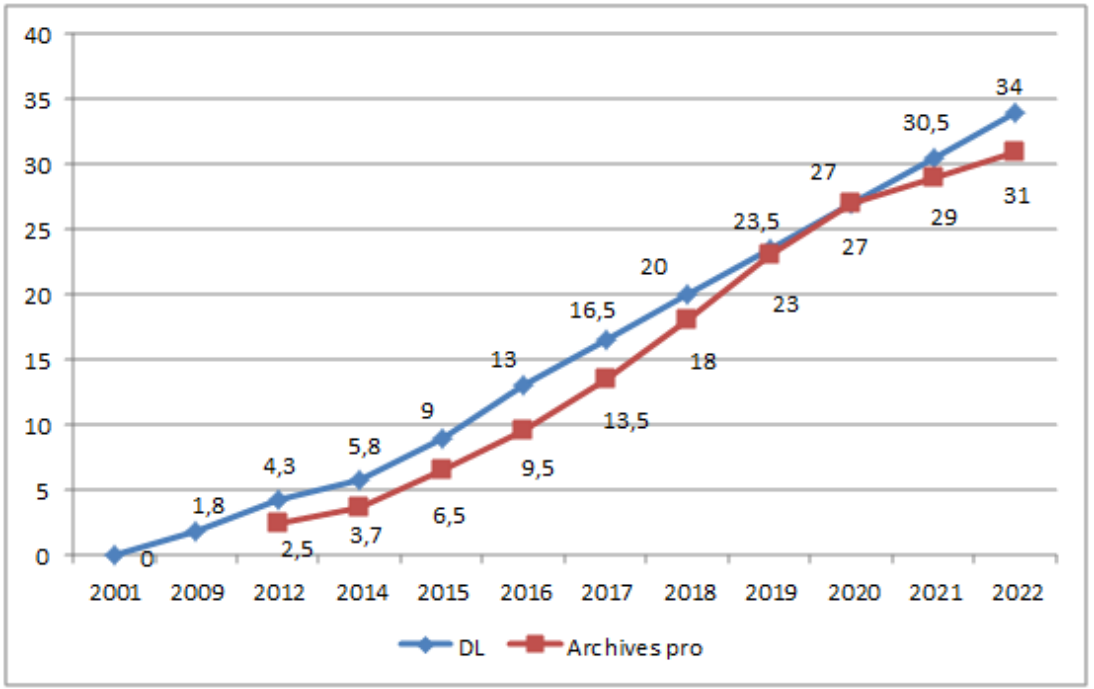
Comment gérer la volumétrie et le stockage ?

Une production de plus en plus importante :

- Taille des fichiers générés (images au format TIFF, fichiers audiovisuels)

Des questions de coût, mais pas seulement :

- Une sauvegarde devenue impossible
- Des migrations régulières rendues indispensables
- Revoir les prescriptions en matière de production, pour optimiser les coûts de stockage (cf. réflexion de la BnF et de l'INA en faveur de l'adoption du Jpeg2000)



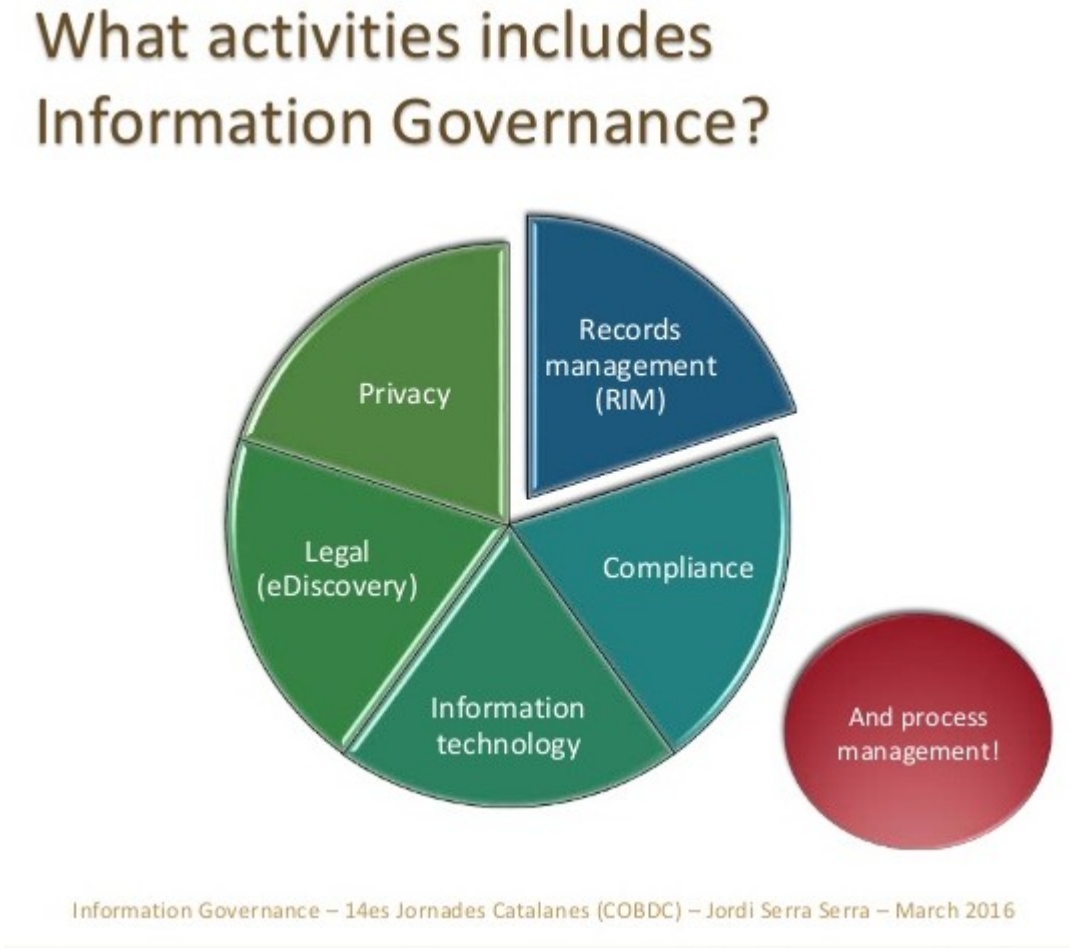
Courbe de l'évolution prévisionnelle (en Po) de l'archivage numérique du dépôt légal et des archives professionnelles

Patrimoine nativement numérique

Enjeux institutionnels/organisationnels - 1

Jusqu'où peut-on aller dans l'intervention sur la création de ce patrimoine ? :

- Gouvernance de l'information et *records management*, une solution ?
- Quelles contraintes sont supportables par les créateurs de patrimoine ?
- Comment interagir avec les professions liées à l'informatique (DSI, RSSI) et à la gestion de la donnée ? (DPD, ADD, data scientists)



Enjeux institutionnels/organisationnels - 2

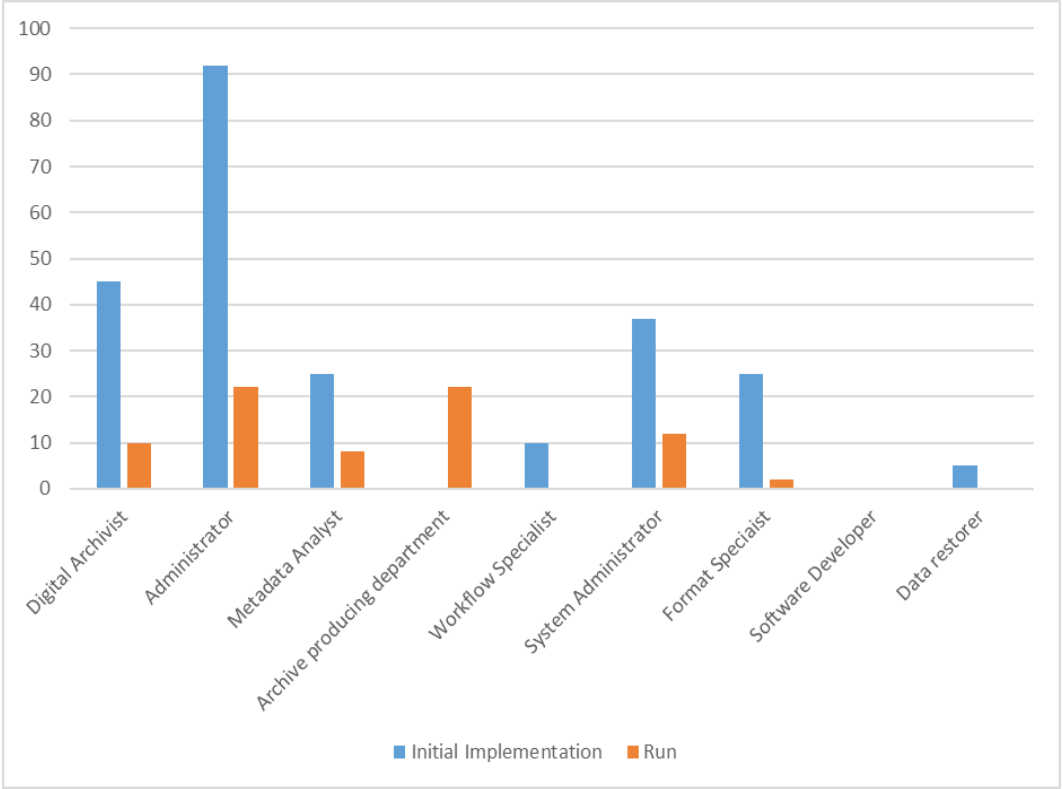
Quelle organisation pour les institutions patrimoniales ?
Quelles compétences ? :

- Préservation numérique et expertise sur les formats de fichiers
- Administration de systèmes
- Architecture et développement de chaînes de traitement informatique
- Développement logiciel
- Opérations quotidiennes

Faut-il faire évoluer l'offre de services ? :

- Les labs

Où positionner compétences et savoir-faire ?



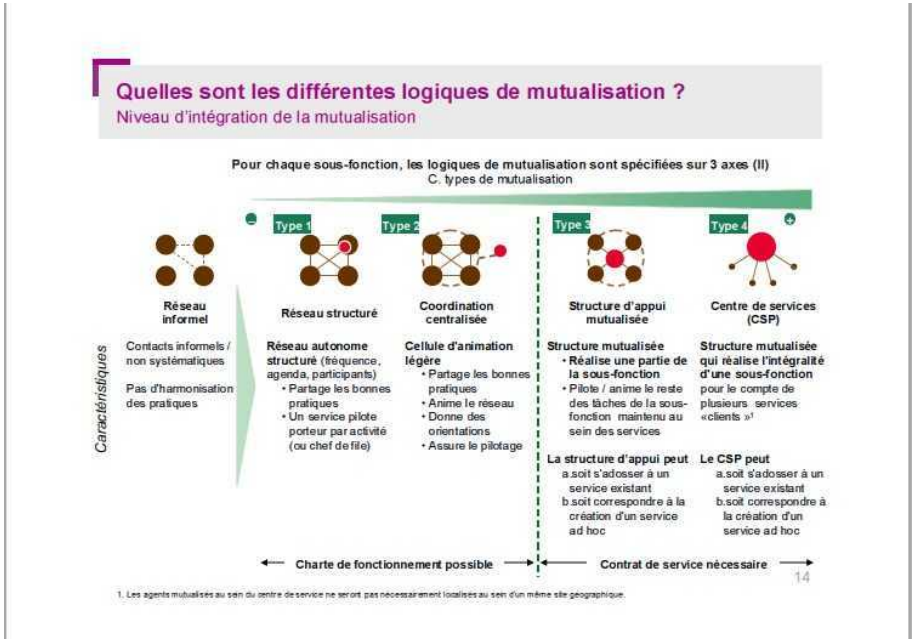
Enjeux institutionnels/organisationnels - 3

Comment adapter le cadre théorique et prztiques à l'évolution de la création ? :

- Centralisation de la production des archives dans des systèmes nationaux => vers un archivage au niveau national d'archives auparavant prises en charge au niveau local
- Mutualisation interdépartementale

Conséquences :

- Une gestion des archives à l'ère post-custodiale : importance de l'accès et de la répartition des responsabilités
- Importance du développement des services d'accès



Enjeux institutionnels/organisationnels - 4

Comment passer à l'échelle et faire face au défi de la massification ? L'exemple de l'INA

Une double massification :

- Massification en terme de périmètre
- Massification en terme de volumétrie

Quelles conséquences ? :

- Faut-il prioriser ? Si oui, quoi ?
- Comment industrialiser les processus de collecte ? Les processus de traitement ? Les processus de mise à disposition ?
- Avec quels financements et quelles conséquences sur les pratiques métier ?
- Avec quels outils et quelles compétences ?
- Quelle conduite du changement ?

Enjeux institutionnels/organisationnels - 5

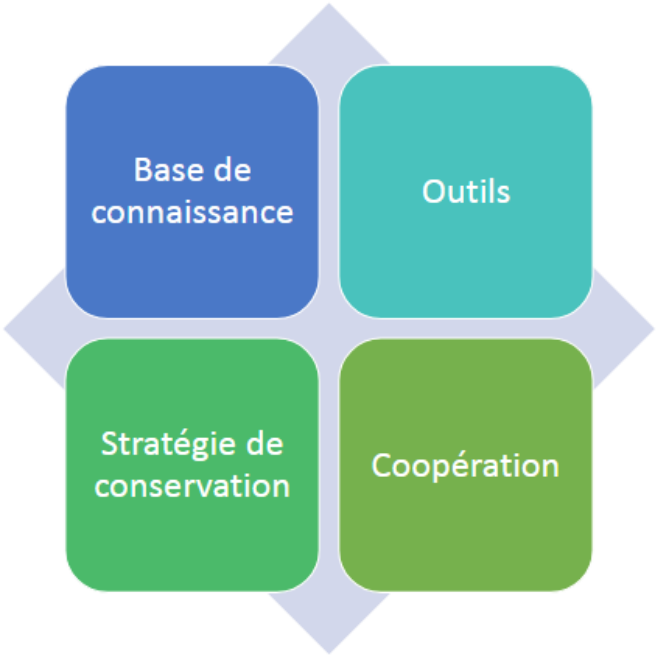
La difficulté à organiser la recherche en préservation du numérique natif

- De nombreuses thématiques : processus métiers, compréhension du parimoine nativement numérique (diplomatique, sciences de l'information), préservation numérique, place de l'intelligence artificielle
- Mais une difficulté à structurer et à financers des programmes de recherche (ex. Balise)

La nécessité de développer et de structurer une veille technologique partagée :

- Des domaines très spécialisés et très techniques : formats de fichiers (évaluation, identification, validation, etc.), émulation, migration de format, accès
- Comment répartir les efforts ?

L'enjeu des communautés non anglophones



Enjeux historiques/scientifiques - 1

Comprendre la genèse du corpus : exemple de la collecte des sites web – la BnF

Qu'est-ce qui est collecté ?

- Collecte « large », réalisée une fois par an, des sites français communiqués par l'Afnic et OVH (0,9 million de sites en 2007, 4,5 millions de site en 2017)
- Collecte « ciblée » :
  - Des sites sélectionnés par des partenaires et les biliothécaires de la BnF
  - Une fréquence de collecte adaptée
  - Distinction entre :
    - Les collectes « courantes » portant sur des sites de référence
    - Les collectes « projets » sur des thématiques transverses ou des événements majeurs (ex. : commémoration du centenaire de la Grande Guerre)
    - Les collectes « d'urgence » pour des événements inattendus (ex. : Covid)

Comment est-ce collecté ?

- Collecte automatisée à l'aide d'un robot-logiciel qui explore les sites
- Profondeur variable en fonction du site
- Chaque collecte est datée

Enjeux historiques/scientifiques - 2

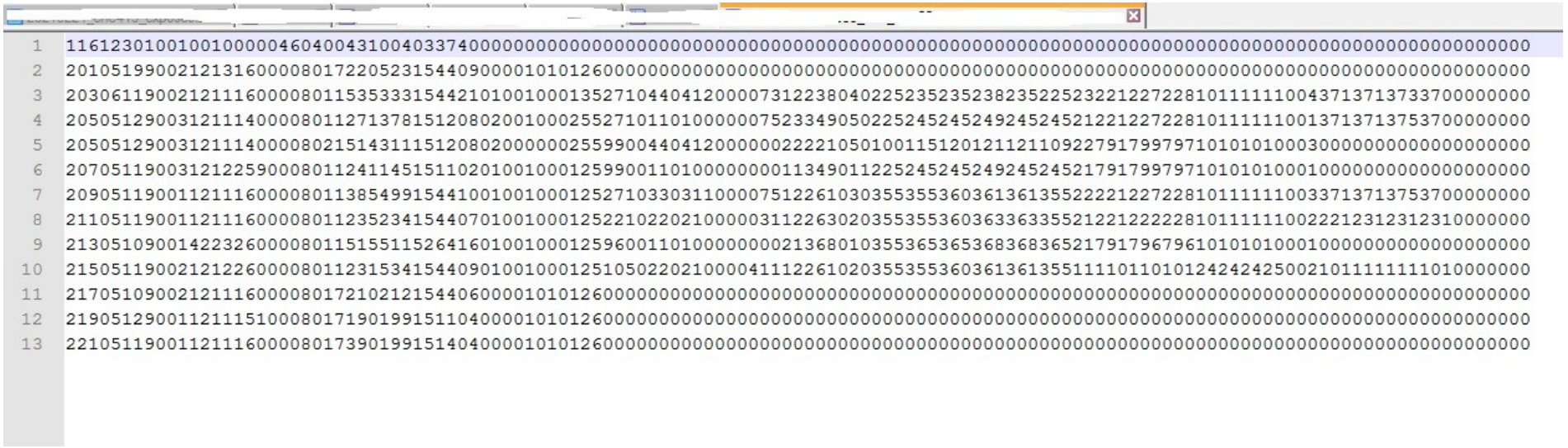
Comprendre la constitution du corpus. Exemple de la collecte des informations des comptes Tweeter est dépendante de la source utilisée

Modalité de collecte	Données			Structure		
	Visibles	Métadonnées	Contexte	Graphique	Interaction	Informatique
Capture d'écran	Oui (format image)	Non	Non	Oui (pour un terminal)	Non	Non
Aspiration web	Oui	Non	Oui (seulement visibles)	Oui (pour un terminal)	Oui (pour un terminal)	Non
Aspiration web récursive	Oui	Non	Oui (seulement visibles)	Oui (pour un terminal)	Oui (pour un terminal)	Non
Requêtes API	Oui (images exclues)	Oui	Non	Non	Non	Oui (JSON)
Requêtes API récursives	Oui	Oui	Oui	Non	Non	Oui (JSON)

Éléments capturés par différentes modalités de collecte des tweets
Source : Antonin Segault, « Documenter Twitter : défis et méthodes pour la constitution de corpus de tweets », Balisages [En ligne], 1 | 2020, mis en ligne le 24 février 2020, consulté le 25 mars 2021. URL : https://publications-prairial.fr/balisages/index.php?id=280

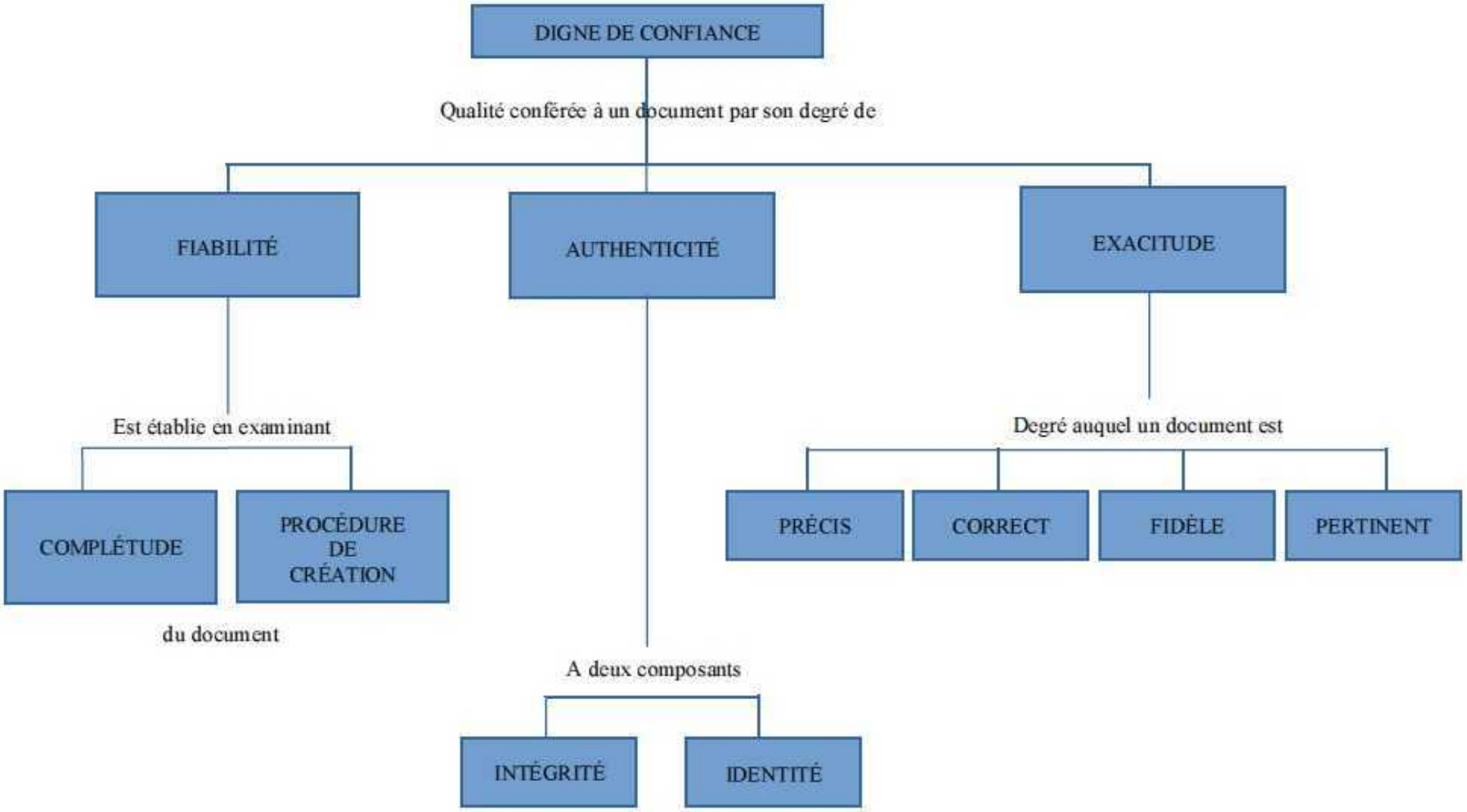
Enjeux historiques/scientifiques - 3

Comprendre le corpus (genèse, tradition), mais aussi son contenu



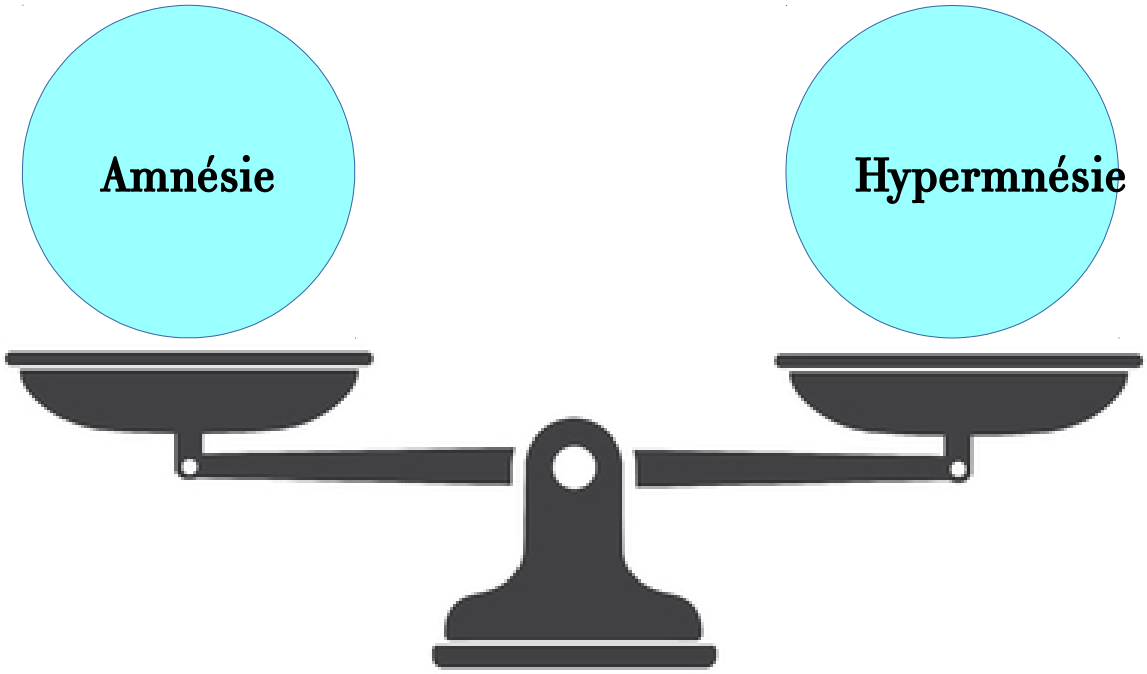
Enjeux historiques/scientifiques - 4

Avoir confiance dans le contenu d'un corpus



Enjeux mémoriels/sociétaux - 1

Jusqu'où aller dans l'acquisition du patrimoine nativement numérique ?



Enjeux mémoriels/sociétaux - 2

Comment conserver et gérer la mémoire individuelle/familiale ?

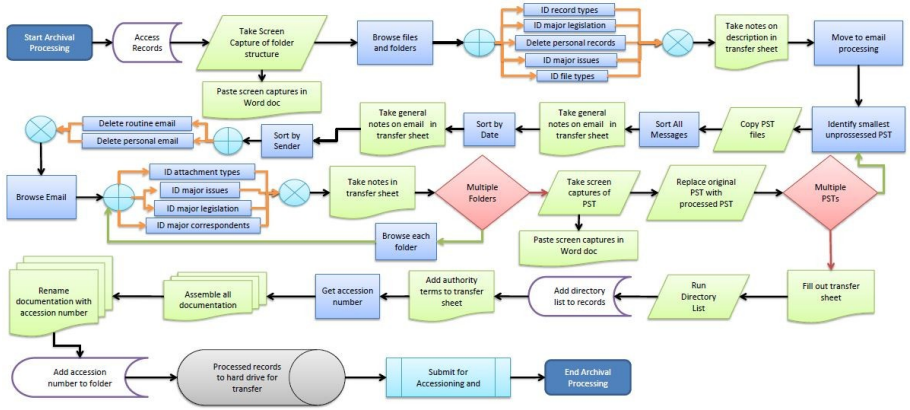
Enjeux techniques - 1

Comment prévenir l'obsolescence des supports de stockage et des formats de fichiers ? :

- Définition d'une politique de préservation
- Contrôle et surveillance des environnements
- Capacité à qualifier le stock conservé ou à prendre en charge et à identifier les éventuelles risques à gérer

Mettre en œuvre une chaîne de pré-traitement et de traitement, acceptable par les créateurs :

- Captation vs. Extraction
- Capacité à mettre en œuvre des opérations d'extraction de contenus (forensics)
- Définition et mise en œuvre des flux de données, en gérant les questions de sécurité et les ruptures de chaîne



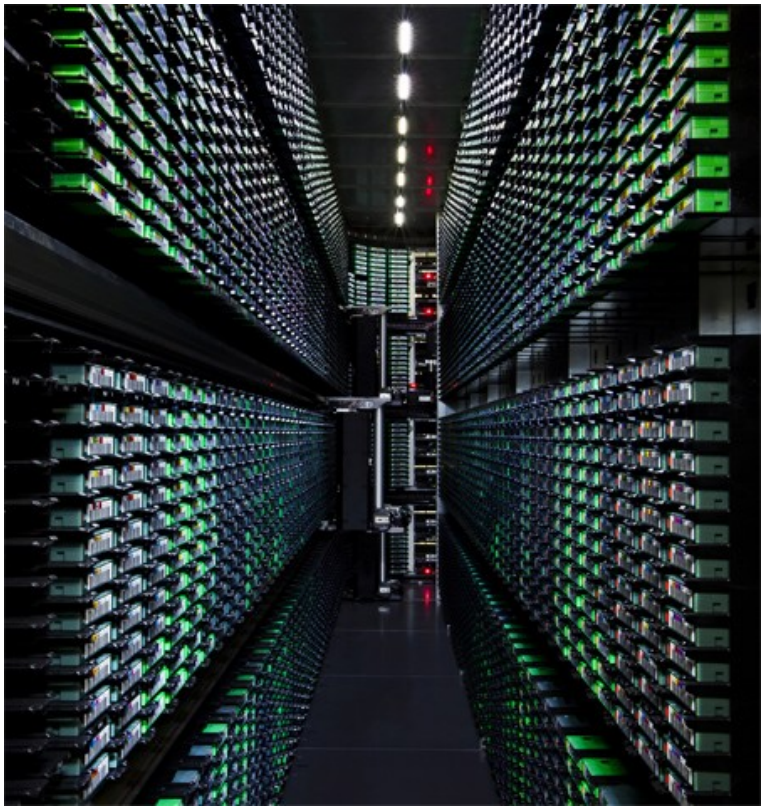
Enjeux techniques - 2

S'assurer de la fiabilité, de la sécurité et de l'évolutivité des plateformes d'archivage

- Robustesse de l'environnement
- Bonne prise en compte des questions de stockage (sauvegarde vs. réplication, différenciation des offres de stockage, migrations régulières), et de la surveillance des infrastructures de stockage
- Existence de bons mécanismes d'audit
- Traçabilité des actions
- Capacité à gérer un PRA/PCA
- Prise en compte des problématiques de sécurité

Disposer de la capacité de calcul permettant de faire les traitements attendus :

- Prise en charge d'enregistrement nombreux et de volumétrie importantes et régulières
- Capacité à mettre en œuvre des opérations de traitement : conversion de format, opérations d'intelligence artificielle



Enjeux techniques - 3

Garantir la lisibilité et l'intelligibilité sur le long terme :

- Quelle méthode utiliser ? Migration de format ? Émulation ? Choix de l'interface ?

Envisager modalités et services d'accès :

- Fournir un accès sur site ou un accès à distance, tout en respectant les contraintes légales ?
- Définir les interfaces d'accès
- Services offerts aux usagers (notamment aux chercheurs), y compris en matière d'exploitation des données

Conclusion

Points communs et différences

Points communs

Capacité à gérer les enjeux organisationnels, même si ceux-ci ne sont pas totalement similaires

Capacité à faire comprendre la constitution des corpus

Besoin de gérer des volumes de données croissants et besoin de capacités de traitement croissantes (puissance de calcul)

Tendance à croire que le numérique est sans limites

Enjeux croissants d'automatisation

Différences

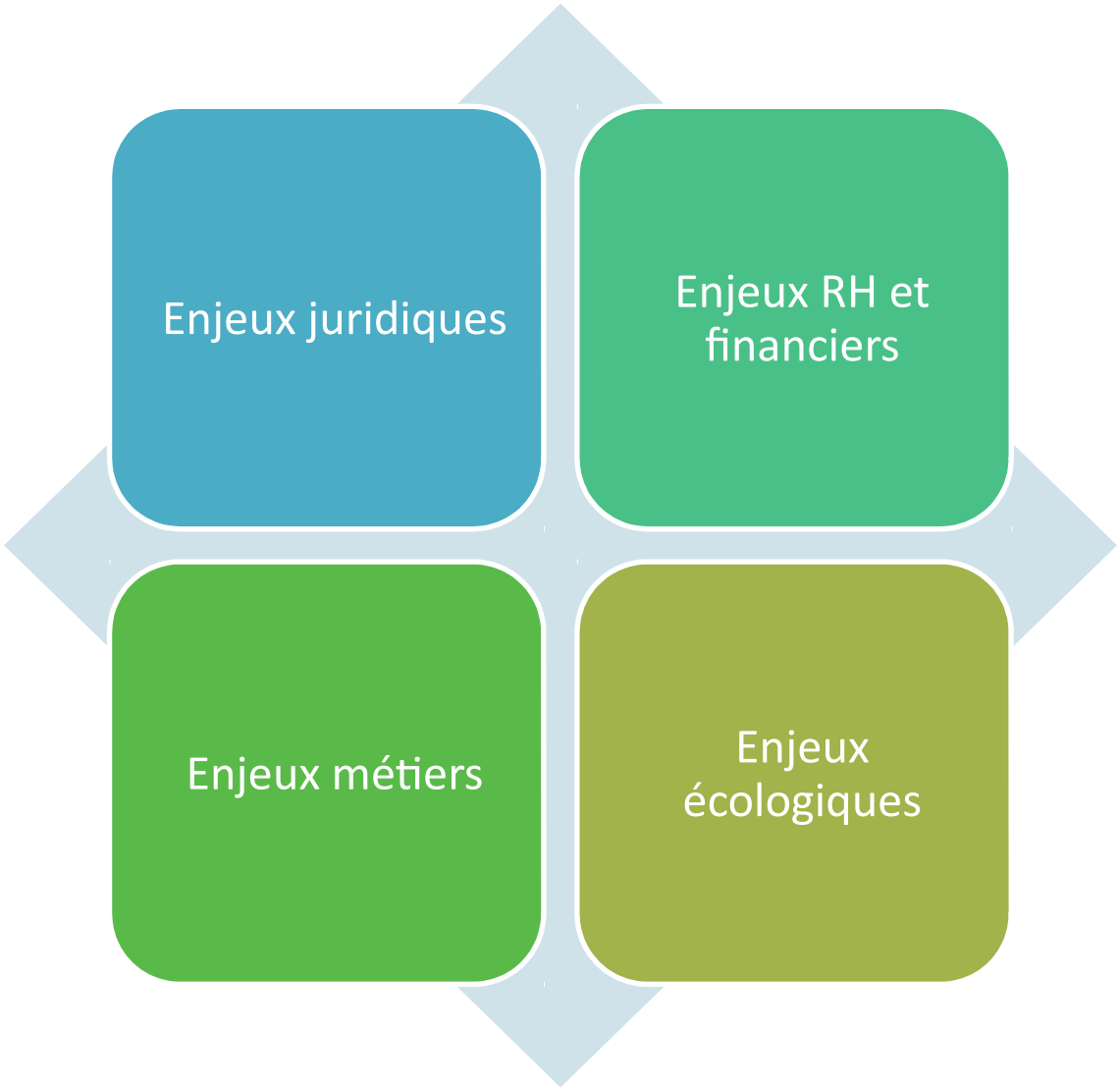
Criticité de la perte de données

Capacité à maîtriser l'environnement de création

Capacité à appréhender la genèse et la tradition des sources, et à garantir la confiance en celles-ci

Enjeux techniques

D’autres enjeux



Discussion

Édouard Vasseur
École nationale des chartes
65, rue de Richelieu
75002 Paris
FRANCE
edouard.vasseur@chartes.psl.eu

École nationale des chartes | PSL