

Les défis de la préservation face au déluge de données numériques

Séminaire de l'Association Aristote



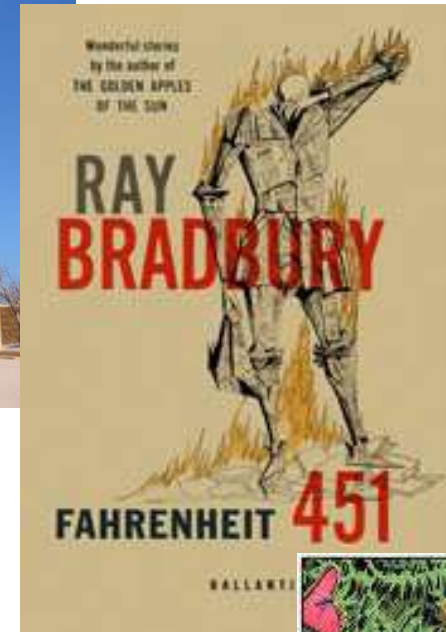
Inflation des données numériques : que faire ? est-ce viable ?



Nicolas Larrousse
IR* Huma-Num UAR 3598
Groupe PIN (Préservation de l'Information Numérique)



La préservation par tradition orale

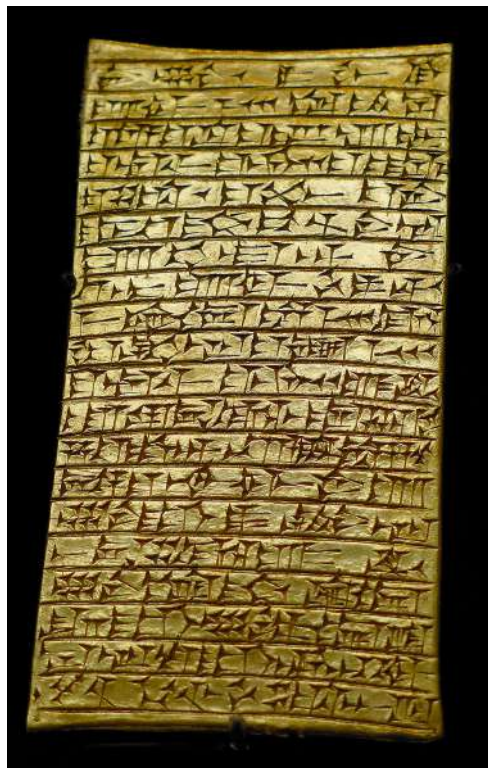


Des arbres à palabres en Afrique à la dystopie de Ray Bradbury où les résistants deviennent des livres pour remplacer ceux brûlés par les pompiers en passant plus récemment par « Préférence Système » de Ugo Bienvenu qui renverse la perspective sur la transmission





La préservation par l'écrit



Du cunéiforme au manuscrits médiévaux des moines copistes en passant par les glyphes des Mayas ... et évidemment l'imprimerie





La préservation par le mélange de traditions orales et écrites



La musique se transmet par les partitions mais la mémoire de son interprétation se fait (notamment) via les conservatoires



Le binaire ... codage ultime ?

TABLE SE MEMOIRES DE L'ACADEMIE ROYALE
DES
NOMBRES

Les chiffres au-dessus du double du plus haut degré. Car icy, c'est comme si on dit, par exemple, que 111 est la somme de quatre, de deux & d'un. Et que 101 ou 1 est la somme de huit, quatre & un. Cette propriété les aux filiales pour peindre toutes les unités avec peu de poids, & pour servir dans les monnoies pour donner plusieurs valeurs avec peu de pièces.

Cette expression des Nombres étant établie, il est à faire très-facilement toutes sortes d'opérations.

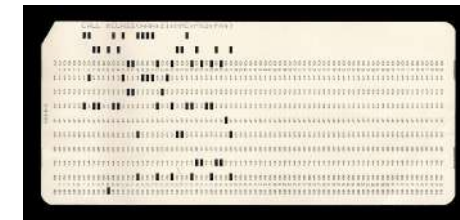
Pour l'Addition par exemple.

Pour la Soustraction.

Pour la Multiplication.

Pour la Division.

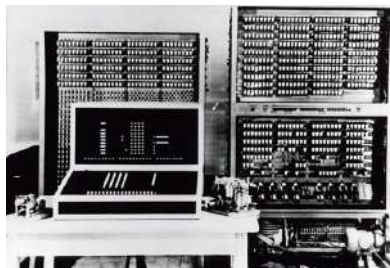
Et toutes ces opérations font si aisées, qu'on n'a jamais besoin de s'en effrayer ni deviner, comme il faut faire dans le calcul ordinaire. On n'a point besoin non plus de s'en apprendre par cœur icy, comme il faut faire dans le calcul ordinaire, où il faut savoir, par exemple, que 6 & 7 pris ensemble font 13, & que 3 multiplié par 3 donne 9, suivant la Table d'une fois au-dessus, qu'on appelle Pythagoriques. Mais icy nous cela se trouve de la preuve de l'histoire, comme l'un voit dans les exemples précédents sous les figures 3 & 6.



```

<measure n="5">
  <staff n="1">
    <layer n="1">
      <note pname="c" oct="5" dur="8" dots="1" stem.dir="down" accid="n">
        <vervec n="2">
          <syll wordpos="1" con="d">mo</syll>
        </vervec>
      </note>
      <note pname="b" oct="4" dur="16" stem.dir="down">
        <vervec n="1">
          <syll wordpos="2" con="n">na</syll>
        </vervec>
      </note>
      <note pname="b" oct="4" dur="16" stem.dir="up">
        <vervec n="1">
          <syll wordpos="3" con="d">sch</syll>
        </vervec>
      </note>
      <note pname="d" oct="5" dur="16" stem.dir="up">
        <vervec n="1">
          <syll wordpos="4" con="t">ten</syll>
        </vervec>
      </note>
    </layer>
  </staff>
</measure>

```



Depuis Leibniz pour le calcul en passant par les métiers Jacquard et les cartes perforées au stockage dans le cloud

Indépendance totale par rapport au support employé



Le binaire ... vecteur d'amnésie ultime ?

Perdre l'accès à ses données numériques est (très) facile !

Pour pouvoir accéder à une donnée numérique il faut être capable de maintenir un ensemble de technologies intriquées

Les risques sont donc multiples :

- Obsolescence du matériel de lecture du support ou des composants associés
 - Supports détériorés
 - Format de la donnée illisible
 - Outils d'interprétation indisponible
 - Pas de métadonnées accessibles
- Etc.





Les supports de préservation

Classiquement aujourd'hui, on stocke sur des supports utilisant différentes technologies (disques, bandes ...)

La préservation est « active », elle nécessite de l'énergie et des migrations fréquentes voire permanentes





Les supports de préservation

D'autres technologies plus « passives » sont utilisées, mais plus marginalement



Réserve mondiale de semences du Svalbard



Le stockage sur microfilms

Utilisé dans le projet Artic Vault Code, le code, essentiellement venant de GitHub est exprimé sous forme de QR code (technologie de la société Piql). Le support est auto descripteur.



GitHub Open Source Archive Box
Design by Alex Maki-Jokela



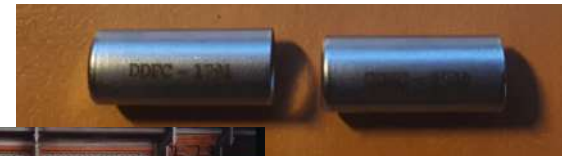
Les supports de préservation

D'autres technologies plus « passives » sont utilisées, mais plus marginalement

Déclaration des droits de l'homme et de la femme stockées sur ADN aux Archives Nationales

Le stockage sur ADN

Technologie intéressante et prometteuse , mais les vitesses d'écriture sont encore trop faible



» Armoire de fer »
des Archives
Nationales

Le stockage sur verre

Le projet Silica
de Microsoft





Les formats de données

Le format dans lequel est exprimé une donnée pose les questions de son identification et également de l'outillage nécessaire pour le lire. Il est donc nécessaire de combiner ces deux aspects pour être capable de préserver la lisibilité .

Matrice de risque des Archives Américaines

Les formats sont classés selon différents critères avec une évaluation du risque et une pondération

15,00	Moderate Risk	NF00545	Cold Fusion Component File
23,00	Low Risk	NF00142	Cold Fusion Markup Language
26,00	Low Risk	NF00143	Comma Separated Values
21,00	Moderate Risk	NF00725	Command manual
19,00	Moderate Risk	NF00144	Common Data Format Toolkit
-24,00	High Risk	NF00790	CompeGPS Land Track file
-13,00	Moderate Risk	NF00111	Compressed Archive File
-41,00	High Risk	NF00774	Compressed ARX file archive
-5,00	Moderate Risk	NF00691	Compressed Dynamic Link Library (DLL)
15,00	Moderate Risk	NF00145	Computer Graphics Metafile Binary 1
15,00	Moderate Risk	NF00546	Computer Graphics Metafile Binary 2
15,00	Moderate Risk	NF00547	Computer Graphics Metafile Binary 3
15,00	Moderate Risk	NF00548	Computer Graphics Metafile Binary 4
15,00	Moderate Risk	NF00783	Computer Graphics Metafile Binary unspecified version
-1,00	Moderate Risk	NF00146	Configuration File
-24,00	High Risk	NF00148	Corel CMX Compressed

NARA Risk Matrix



Les formats de données

Le format dans lequel est exprimé une donnée pose les questions de son identification et également de l'outillage nécessaire pour le lire. Il est donc nécessaire de combiner ces deux aspects pour être capable de préserver la lisibilité .

Identification des formats

Le projet PRONOM des Archives Britanniques de liste de référence sur les formats intègre un outil, DROID, qui permet une identification fine



The **technical registry**
PRONOM

DROID



Travail général sur les formats

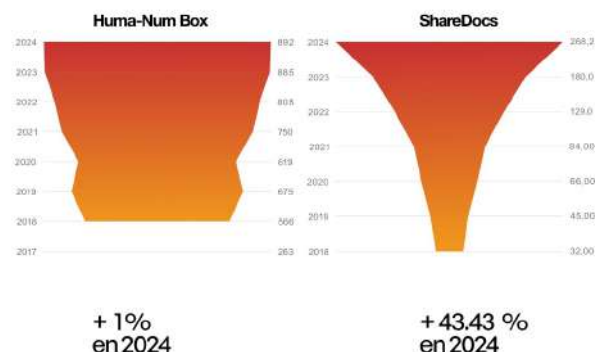
Le groupe PIN anime un groupe national consacrés aux formats



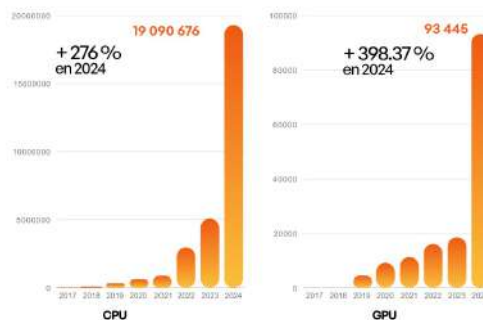
Le « déluge » de données en SHS

En quelques années, les modalités de constitution de corpus en Sciences Humaines et Sociales ont fortement évolué : cela est dû notamment à la mise en oeuvre de nouveaux dispositifs comme ceux de transcription automatique, de reconnaissance d'écriture manuscrite mais aussi de l'apparition de nouvelles technologies comme le LIDAR.

Stockage Volume de données stockées (en To)

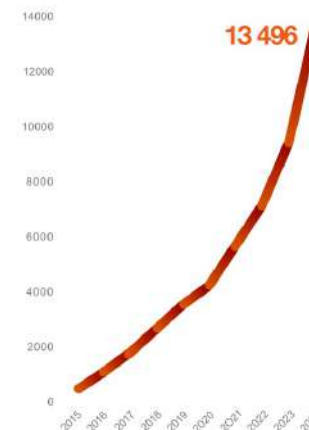


Traitement Heures consommées en HSO6



Nombre de comptes ShareDocs

+ 41.23 %
soit **3 940** comptes
créés en 2024



Des approches méthodologiques nouvelles ainsi que des modes de représentation adaptés qui conduisent naturellement au développement de nouveaux outils.



Le « déluge » de données en SHS

La préservation de ces données pose de nombreuses questions :

- Que doit on préserver et sur quels critères ?
- Combien de temps doit-on conserver ces données ?
- Est-ce que l'on impose de formats, des métadonnées ?
- Les données de la recherche sont destinées
« administrativement » à être versées aux archives nationales.
Lesquelles doivent être proposées ?

Etc.



Expérience acquise depuis 10 ans avec le dépôt de données au CINES de certains jeux de données, mais pas de solution globale

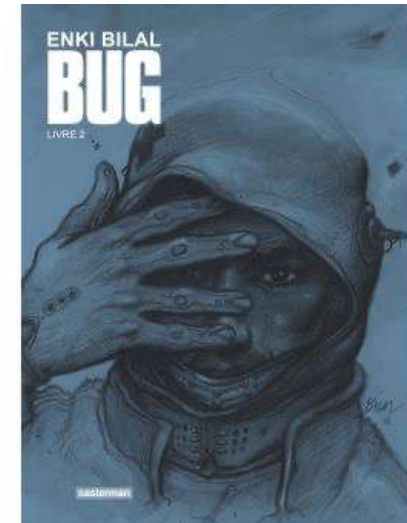


Le binaire ... vecteur d'amnésie ultime ?

Perdre des objets numériques est facile !

La préservation n'est plus totalement passive !!
Nécessité de préserver aussi la « lutherie »,
l'information de codage et les métadonnées.
Il est utile d'y consacrer des ressources
importantes.

Que se passera t-il en cas de « panne générale »



Nécessité de « trier », de construire des règles etc. pour la préservation

Dans l'ouvrage « Préférence Système » (Ugo Bienvenu), un « comité » se réunit pour décider de la conservation de données



Le binaire ... vecteur d'amnésie ultime ?

La peste de l'oubli



Cent ans de solitude G. Garcia-Marquez

La perte de mémoire oblige les habitants à inventer une méthode pour se souvenir des choses et Aureliano commence à étiqueter tous les objets pour se rappeler leur nom ; cependant, cette méthode commence à faillir quand les personnages oublient la lecture. Jusqu'au jour où revient Melquíades avec une boisson pour rétablir la mémoire de manière immédiate.