aristote

À la croisée des révolutions numériques

Bibliothèque Nationale de France
Site François Mitterand (BNF)
Paris

Les donnes numériques, une espèce en voie d'extinction ?

Jeudi 2 décembre 2021

Coordination scientifique :

- **Françoise Banat-Berger** (Ministère de la culture),
- **Laurent Duplouy** (BnF),
- **Céline Guyon** (AAF),
- **Nicolas Larrousse** (Huma-Num),
- **Olivier Rouchon** (Cines).

association des archivistes français

Huma-Num

CINES
Centre Informatique National de l'Enseignement Supérieur

BnF Bibliothèque nationale de France

Renseignements, programme…
https://www.association-aristote.fr/internet-quantique/

Compte Rendu du séminaire du 2 décembre 2021

« LES DONNEES NUMERIQUES, UNE ESPECE EN VOIE D’EXTINCTION ? »

Grand Auditorium de la Bibliothèque Nationale de France

Organisé par le groupe PIN (Aristote)

Table des matières

Ouverture

Christophe Calvin, président de l'association Aristote, et Céline Guyon, présidente de l'Association des Archivistes Français (AAF).

Christophe Clavin rappelle que le groupe PIN est un des plus vieux groupes de réflexion d'Aristote, qui ne s'est jamais arrêté de travailler, et constitue une des bases de l'association. Il introduit alors le sujet : les données numériques étant de plus en plus nombreuses, et donc, en nombre, très loin d'être en "voie d'extinction", comment les archiver de manière ordonnée, afin qu'elles soient utilisées convenablement.
Céline Guyon, président de l'Association des Archivistes Français (AAF) rappelle le rôle des archivistes est de collecter et stocker les données numériques pour une durée indéterminée - le point important - et les rendre accessibles et organisées pour les générations futures. Elle rappelle qu'elle a commencé à s'intéresser à la question de la pérennisation des données à l'occasion d'un stage organisé par le groupe PIN. Elle souligne la joie d'être associée à l'association Aristote, pour son regard pluridisciplinaire, car la confrontation des regards est essentiel, au-delà de nos communautés respectives pour aborder les enjeux de la pérennisation sous ses multiples dimensions.

Entretien vidéo de David Giaretta

David Giaretta, directeur du PTAB

VIdéo ici

David Giaretta revient à la source de la définition d'un document. Après avoir rappelé l'étendu de la croissance des données dans le monde, il précise que des bits ne sont qu'une manière de stockées les informations, et plus globalement, qu'elle que soit la méthode avec laquelle ce stockage a lieu, l'information peut signifier à peu près tout ce qu'on veut.
Il rappelle qu'il existe de nombreuses menaces sur la préservation des données. Et prévient qu'il faut faire attention avec le numérique : les utilisateurs pourraient ne pas comprendre les phrases, le format ou les algorithmes impliqués. Il souligne le manque généralisé de hardware disponible en l'état, de logiciel de support, ou d'environnement informatique. Avec comme conséquences, que les preuves pourraient être perdues, parce que dans ce cas, l'origine ou l'authenticité ne peut être prouvée. Les restrictions d'accès pourraient aussi ne pas être respectées. Un autre danger serait la perte de la capacité à localiser la donnée. De savoir qu'elle existe mais être incapable de la trouver.

David Giaretta revient sur les liens entre les standards et les OAIS (Open Archival Information System) et pose la question d'un archivage de confiance. Qu'est-ce exactement ? sur quoi porte alors la confiance ? Ce sont encore là des choses à mieux définir.

Mais un des plus grands dangers serait d'oublier que l'archivage se prépare et se réalise au présent. Qu'elle que soit la méthode. Un danger à bien avoir en tête, car il questionne la valeur de l'information à apprécier, le besoin de contexte ou non, et si oui, lequel. Auquel cas n'importe quelle archive pourrait devenir une mauvaise ressource.

La préservation numérique, historique et perspectives vues au prisme du groupe PIN

Olivier Rouchon, directeur du département archivage et diffusion depuis 2009, au CINES Centre Informatique National de l'Enseignement Supérieur.

Vidéo ici

Olivier Rouchon est le fondateur du groupe PIN, qu'il a porté pendant de nombreuses années. Il veut présenter comment PIN a accompagné la communauté des archivistes. Il rappelle que ce sont les 20 ans du groupe, créé en 2000. Il revient sur le mode de fonctionnement du groupe (réunions, forum de discussion, site Web, groupe Linkedin, serveurs web en cours de construction…) Les objectifs du groupe : traiter toutes les questions relatives à la pérennisation et préservation et de l'information sous forme numérique, que ce soit technique, ou juridique et organisationnel. Au total PIN a réalisé, 55 plénières sur 19 ans. Le panel d'intervenants a été très variés. En moyenne les réunions ont compté 35 participants, ont vu 258 présentations et 83 organismes différents ont participé. 2009 a vu un tournant dans l'organisation, avec des animations davantage tournantes de l'organisation des plénières.

Olivier Rouchon revient ensuite sur les thématiques abordées, qu'il classe en cinq catégories :

La question des normes références, les applications et outils, la technologie, les questions liées aux organisations, les retours d'expérience.

Il a évalué la récurrence de ces thématiques au cours du temps. Résultat : petit à petit sur les vingt dernières années, on parle de moins en moins des normes et des références, et de plus en plus des retours d'expérience. Les normes étant arrivées tardivement - la norme ISO ne date que 2002 - il y avait un gros travail de pédagogie à faire au départ. Puis à partir de 2005, il y a une montée ne maturité de la communauté. Les standards sont assimilés dans la communauté, et les membres veulent les implémenter. Il effectue le même travail au niveau des sous-groupes de travail et des formations.

Il revient ensuite sur les différentes publications qui ont été réalisées et questionnent la suite de certains sujets à l'heure actuelle. Enfin il revient sur les perspectives du groupe PIN, et demande notamment à l'audience de proposer des sujets qui pourraient la concerner.

De la carte perforée aux données : une plongée rétrospective dans la pratique de l'archivage électronique

Céline Guyon, présidente de l'AAF

Vidéo ici

Céline Guyon rappelle que l'AAF avait créé un groupe en 1967 institué : Mécanographie électronique, et s'intéressait aux intrications entre archives et informatique. Elle se base sur un article de 1971 paru dans la gazette des archives, une revue révélatrice sur l'influence du numérique sur archivage. Selon elle, l'archivage a été modifié en profondeur. Le milieu des archivistes a vu se multiplier toutes les techniques de reproduction comme jamais dans l'histoire (bande magnétique, carte perforée, fichier informatique, copie carbone). L'amas de données a explosé. Il a fallu redéfinir l'objet archive, et par conséquent, toute la profession. Elle s'interroge alors : nos questionnements actuels sont-ils les mêmes ? Certaines questions du passé font-elles écho ? Et si oui, lesquelles sont les

mêmes ? L'article de 1971 questionne ainsi l’impact des cartes perforées sur les pratiques archivistiques. Pour les archivistes, la question était de savoir si la conservation des cartes perforées pouvait se substituer à l’archivage des documents papiers de base. Elle revient alors à travers l'histoire sur les questionnements qui ont égrainés les archivistes dans leur rapport à l'informatique. Pour conclure sur le défi prochain du monde archivistique : il ne sera pas tant lié à la pérennisation des contenus qu'à la question de la navigation à travers ces contenus. L'approche est encore très documentaire, aujourd'hui, dans la documentation administrative. Donc sans navigation particulièrement adaptée.
Elle conclut alors en citant Jean Favier, qui explicite le fait qu'au-delà du support, le document informatique évolue chaque jour. Le document archivé ne rend donc pas compte des étapes successives de production de l'archive. "L'archiviste pour rendre à l'histoire les services qu'on attend de lui devra donc intervenir pendant le fonctionnement, et donc pendant l'évolution du système d'information", cite-t-elle.

L’archiviste au milieu du gué : pourquoi la collecte des archives numériques reste-t-elle si difficile ?

Stéphanie Roussel, pour la société Mintika

La vidéo est ici

Stéphanie Roussel a travaillé dans le monde des archives administratives publiques essentiellement, avant de créer sa société de conseil. Elle a accompagné au total une soixantaine de projet, et nous livre ainsi la synthèse de son expérience. Elle introduit sa présentation en réaffirmant la thèse selon laquelle l'informatique, au-delà de l'objet d'archives change également la manière de travailler de l'archiviste, mais ajoute que le milieu a atteint une certaine forme de maturité sur ce sujet. Selon elle, l'étape de la collecte des données, pourtant étape cruciale du processus d'archivage a été oubliée dans les pratiques, et le secteur de l'archivage fait ainsi face de manière frontale à cette question. Si les outils sont pourtant là la réalité du terrain est tout autre. Sur 8 Téraoctets de collecté (ce qui est assez peu), selon le rapport annuel de l'AAF, les disparités selon les territoires sont énormes : moins d'un quart des départements concentrent l'essentiel des collectes.
Or, les standards existent, et les outils sont là aussi. Pourquoi alors la collecte peut sembler trop faible ou problématique ? Un des premiers problèmes consiste dans la sélection. Les données augmentent de 60% par an, les expansions d'espace de stockage permettent une progression de 25% par an, et les budgets des services d'archives augmentent de 2% par an… D'autre part, les opérations de collecte sont plus complexes aujourd'hui dans le numérique, que dans le monde physique. Elle pose alors la question de l'expertise, souvent dédiée au numérique, et non pas à l'archivage. Stéphanie Roussel aborde alors le fait qu'elle considère qu'il faut passer d'une collecte passive à une collecte actives des données, avec toutes les questions et analyse des risques que cela comporte. Pour conclure, elle affirme que si les archivistes ont tous les moyens techniques et les outils pour y parvenir, il manque surtout de la méthode et de la pratique. Rien n'est perdu, mais rien n'est gagné, l'archiviste est encore au milieu du gué, sur ce sujet.

Une cellule Nationale de veille sur les formats de fichiers. Pour une expertise nationale sur les formats de fichiers.

Lorène Béchard, archiviste pour le Cines

Bertrand Caron, BNF

Lorène Béchard et Bertrand Caron présentent les travaux menés par la cellule de veille sur les formats du groupe PIN. Née en 2010, la cellule est née de la volonté de mutualiser les connaissances sur les formats de fichier et aider la communauté des archivistes qui s'interrogeaient sur l'archivage électronique. La cellule regroupe une dizaine de partenaires afin de réfléchir sur les outils, et de rayonner aussi sur les travaux à l'international. C'est ainsi une vitrine de la réflexion française en matière d'archivage de données. La cellule est composée de quatre groupes de travail principaux : La connaissance des formats, les expertises, la traduction, les outils et corpus. Pour les expertises, les équipes ont réalisés un annuaire de l'expertise format en France en 2020. La groupe connaissance des formats un document regroupant et détaillant les neuf critères à prendre en compte afin de définir une politique de formats. Pour Outil et Corpus, ils ont réalisé un corpus des des fichiers disponibles en ligne, ainsi qu'une catégorisation des outils par type d'opération. Enfin pour le sous-groupe traduction trois documents ont été réalisés : un manuel de préservation numérique, une grille d'évaluation des niveaux de préservation et une version rapide de cette même grille. A ces livrables, quatre webinaires ont été réalisés. Bertrand Caron présent alors les projets de la cellule pour les mois à venir, par groupe de travail.

La reprise de données numériques archivées entre 1983 et 2018 aux Archives Nationales

Émeline Levasseur, cheffe de projet archivage électronique aux Archives Nationales

En Video

Émeline Levasseur fait un retour d'expérience sur le chantier de reprise des données entamé aux Archives Nationales, suite à une nouvelle plateforme d'archivage électronique mise en place en 2018 (Constance), avec un focus particulier sur les enquêtes statistiques et leurs métadonnées.
Les enquêtes statistiques représentent presque la moitié des données versées aux archives depuis 1983 mais en termes de stockage, elles ne représentent que 150 giga sur les 73 Téraoctets de données. Ce sont les premières données dont l'archivage a été industrialisé. Émeline Levasseur décrit ensuite les caractéristiques des enquêtes statistiques et le rôle important que joue leur archivage. Concrètement, le but est de pérenniser les fichiers de données sur des bandes LTO (stockage à froid), à plat, et ils sont nommés de manière homogène. Ces données représentent les réponses aux questionnaires, sont ncodées en ASCII (American Standard Code pour Information Interchange, norme d’encodage des caractères) selon le principe 1 donnée = 1 à n caractères, et il y a autant d’octets que de caractères. Elles sont ensuite archivées à plat dans des documents de type texte et stockées sur bande LTO et les fichiers de données sont gérés et décrits dans une base documentaire. La signification donnée, elle, est conservées via les métadonnées (fiches d’application (information sur l’application versée), dictionnaires de données, fiches de structure, dictionnaires de codes) et la documentation papier associée (instructions aux enquêteurs, questionnaires d’enquête vierges, bilan de l’enquête, publications de la recherche).
La spécialiste, après avoir fourni des exemples de données, revient sur le processus de transfert des archives vers le numérique : rendre les données et métadonnées accessibles, numériser la documentation. Pour cela, bien définir les objectifs (l'existant et la cible) et le périmètre. Cela représente aujourd'hui 150 versements pour 134 applications, un peu plus de 5 000 fichiers de données soit 150 Go, 2 065 fichiers correspondant à la documentation associée en PDF soit 116 Go (travail non fini).

Elle détaille alors le processus d'automatisation des transferts (modélisation des paquets, développement et construction des SIPO), et la constitution de nouvelles données dans la nouvelle plateforme. Le repris automatisé a été effectuée avec le cabinet Mintika. Elle montre alors le résultat de cette opération via un exemple consultable d'une enquête sur les actifs financiers des ménages de 1992. Le transfert est en cours de septembre 2021, et 86 enquêtes ont été versées. Le projet continuera en 2022, après la dernière numérisation

Des données éternelles ? 20 ans d'expérience du centre de données CDPP au CNES

Danièle Boucon, du Cnes

En Vidéo

Danièle Boucon, au sein du Cnes, travaille particulièrement pour Centre de données de la Physique des Plasma. Le CDPP a des données uniques, non reproductibles, parfois âgées de 45 ans, et qui ont nécessité de lourds investissements pour être obtenues (missions spatiales, moyens de mesures couteux…) Elle revient ensuite sur l'historique de l'organisation du CDPP, ainsi que sur l'évolution du mode de stockage de ces données, des bandothèques dans les années 70, Data Lake en 2023, en passant par les différents robots Storagetek. Le laboratoire est passé de 1 Tera à 35 Peta. Elle décrit ensuite les évolutions technologiques côté serveur archive. Globalement, tous les 8 ans, le service est confronté à une migration du catalogue et un changement de serveur. Elle rappelle donc que pour la pérennisation, les équipes sont dépendantes des évolutions techniques. Elle revient ensuite sur l'évolution des standards, qui avec le temps, se sont adaptés en relation avec les outils pour étendre leurs utilisations. Il a fallu donc adopter ces standards et convertir les données dans ces différents formats standardisés. Sont présentés ensuite différents chiffres sur l'archivage au sein du CDPP, où l'on aperçoit que le jeu de données augmente considérablement en 2021, car les données à hébergées sont de plus en plus lourdes (une tendance de fond. Elle précise alors que les données sont consultées et utilisées partout sur la planète, mais majoritairement en Europe, et détaille alors l'évolution thématique des données. Enfin, la dernière partie concerne davantage les évolutions des outils qui doivent permettre à tous les chercheurs concernés, non seulement de consulter mais aussi de traiter les différentes données conservées : "Les données ne sont rien sans les outils", argue-t-elle et conclut en affirmant que la communauté sera amenée à travailler sur l'homogénéisation des formats.

Elle insiste sur les différents partenaires impliqués dans les projets et de l'importance de leur implication dans les missions. Les prochains défis selon Danièle Boucon seront de favoriser la "découverte" des données, ouvrir l'archive aux outils et élargir la thématique, corréler des ensembles de données.

Le projet OligoArchive : stockage de données numériques basé sur l'ADN

Raja Appuswamy, Eurecom

En Vidéo

Raja Appuswamy présente le projet OligoArchive, une idée de stockage de données basée sur l'ADN. Le problème auquel le monde des archives est confronté réside dans l'obsolescence des moyens de conservations, notamment quand les données sont stockées sur des bandes magnétiques, et que les lecteurs évoluent tous les 5 à 7 ans. Sur de longues périodes, il existe forcément un moment où la compatibilité ne sera plus assurée. Il prend pour exemple le centre des Archives Nationales Danoises, pour des dessins datant du XVIe siècle et relevant d'une importance nationale unique. Ou encore Hollywood, qui sera confronté à une période d'archives "mortes", car la plupart des films des années 90 à 2000, prolifiques en termes de cinéma, ont été archivés sur des bandes magnétiques, et ils ne réussiront pas à tous archiver convenablement ou à assurer leur pérennité. Si les gros de l'industrie entreprennent des travaux, beaucoup de films indépendants risqueraient d'être perdus. "Nous sommes réellement à un point d'inflexion sur cette question, qui sera déterminante pour les générations futures", estime-t-il.

Pour répondre à la question de l'obsolescence du format, le projet OligoArchive veut changer le support. Lequel résiste bien au temps et permet de stocker énormément d'informations ? L'ADN. Concrètement, le procédé consiste à convertir les *donnsées* dans un procédé standard SIARD (Software Independant Archiving of Relationnal Databases), à le convertir en données ACGT (les acides aminés) puis à créer de l'ADN de synthèse, qui contient, en termes d'informations, les données de départ. Raja Appuswamy revient alors sur tout le process de création et de conservation de l'information. Au bout du compte, le projet n'en est qu'à ses début, et dépend intégralement de l'avancée en termes de création d'ADN de synthèse. Ce serait à ce jour, via les avancées technologiques, 10 millions de fois plus cher par quantité de données… Mais il ne désespère pas de son idée prometteuse, qui pourrait résoudre bons nombres de problèmes pour le monde de l'archivage à l'heure actuelle.

Données numériques et dépôt légal : le cas de documents numériques interactifs

Alexandre Wauthier, ingénieur d'étude, BNF

En Vidéo

Avec le développement des oeuvres d'art numérique, s'est petit à petit posé la question de l'archivage numérique de ces oeuvres. Sachant qu'elles se situent à la marge du dépôt légal, à la différence d'oeuvres commercialisés ou éditées en livres, et qu'elles sont interactives. Tout a commencé en 2014, quand le laboratoire Inrev de l'université Paris 8 ont fait dont d'une trentaine d'oeuvres d'art numérique à la BNF, pour qu'elles soient accessibles au grand public. L'accessibilité et le catalogage de ces oeuvres sont alors devenus un enjeu. Alexandre Wauthier présente alors les différents projets qui ont permis de poser les différentes questions relatives à l'archivage de ces oeuvre, pour aboutir quelques années plus tard à cinq axes majeurs : : la conservation, la documentation, le catalogage, l’histoire des technologies et la terminologie, et la publication de divers outils et guide.

Un des enjeux nouveaux qui est apparu concerne la rejouabilité de l'oeuvre, issu de son interactivité. Comment assurer la pérennisation de la rejouabilité.

Ce champ de recherche a été poursuivi au sein du projet Machine à Lire les Arts Numérique, interfaces et médiations, de 2019 à 2021, notamment sur la création d'une interface graphique et d'un dispositif technique pour la recherche et la consultation de tels documents. Le modèle abouti est hybride, avec à la fois des machines d'époque, mais aussi des émulateurs virtuels qui recréent les conditions techniques, pour assurer la pérennisation. Cela permet d'exécuter le document et l'expérience dans son environnement d'origine. Alexandre Wauthier détaille alors le projet, et les différentes solutions qui ont été déterminées tant logiciels que hardware. Les notices d'utilisation sont adjointes et encapsulées avec le projet initial. Se pose alors la question des types d'oeuvres et de leur statut légal propres, ainsi que de leur objectif intrinsèque. Deux catégories ressortent donc: les oeuvres relevant d’un dispositif de médiation de type « exposition », plus adaptés à un environnement muséal, et les oeuvres relevant d’un dispositif de médiation de type « consultation », adaptés aux bibliothèques publiques.

Table Ronde : Quelles perspectives pour le futur de la conservation

Emmanuelle Bermès, ancienne adjointe scientifique et technique au Directeur des services et des réseaux de la BNF
Bruno Bachimont, professeur à l'université technologique de Compiègne
Ugo Bienvenu, auteur du roman graphique "Préférence Système"
En vidéo

La table ronde s'ouvre notamment par les réflexions d'Ugo Bienvenu, auteur de différents ouvrages et réalisateur de nombreux courts-métrages. Mais les réflexions portent notamment sur son dernier roman graphique "préférence système", où l'auteur imagine, dans un monde où les données ne peuvent plus être conservées, un tribunal qui doit décider de ce qui peut l'être ou ne pas l'être. Selon Ugo Bienvenu, nous commençons déjà, à l'heure actuelle, dans la réalité, à devoir faire des choix. Il donne des exemples concrets, comme Myspace, qui face aux coûts croissants du stockage numérique, a dû déterminer quelles photos devaient être supprimées de la base de données ou non.
Pour cela, le monde doit détermine ce qui est de l'ordre de l'oeuvre patrimoniale, et ce qui est de l'ordre du produit. Il note que de plus en plus, les frontières sont flous, avec des oeuvres qui prennent l'enrobage du produit et inversement. Il faudra donc faire un tric. C'est-à-dire choisir ce qui est déterminant pour le futur ou non. Mais est-ce le rôle de l'archiviste de faire ce tri ?
Selon Emmanuelle Bermès, le rôle de l'archiviste à l'heure actuelle, et de la société en générale, est d'assurer aux générations au génération future la possibilité de faire ce choix en toute conscience, de déterminer en connaissance de cause ce qu'elle voudra garder ou non. Mais cela doit-il se faire au détriment de la planète ? C'est tout l'équilibre à trouver.
Les invités insistent sur le fait que dans la lecture des événements au cours du temps, des erreurs se produisent, et il est important de tout faire pour les limiter, afin d'éviter les erreurs d'interprétation. Notamment quand le produit dépasse les sources historiques. Ainsi en va-t-il de la conférence de Valladolid, où le roman, puis le film, ont tiré les traits des personnages, relisant quelques peu l'histoire, et en fournissant une facette inconforme, ou du moins imprécise de a réalité historique telle qu'elle est décrite dans les sources.
L'archivage sur brin d'ADN, comme présenté précédemment lors de la conférence, est une des nouvelles solutions, et Ugo Bienvenu rappelle que la déclaration universelle des droits de l'homme a été inscrite sur des brins d'ADN et offerte au Bureau National des Archives, mais pose la question, outre de la conservation certes extrêmement longue, de la capacité à pouvoir relire, d'un point de vue technologique, dans le futur, ces informations.
Ainsi, de manière générale, sans répondre à la manière avec le "tribunal" doit effectuer des choix entre ce qui doit être conserver et ce qui ne doit pas l'être, la table ronde invite à réfléchir sur les conditions d'exercice de ce choix. Qui sont les réflexions portées par le monde de l'archivage.