MINISTÈRE
DE L'ÉDUCATION
NATIONALE,
DE LA JEUNESSE
ET DES SPORTS

Liberté
Égalité
Fraternité

VOIE TECHNOLOGIQUE

Série STMG : Sciences et technologies du management et de la gestion

2DE | 1RE | TLE

Management, sciences de gestion et numérique

ENSEIGNEMENT SPECIALITE

GESTION DE CONTENUS ET DONNÉES MASSIVES (BIG DATA) ENSEIGNEMENT COMMUN ET ENSEIGNEMENT SPÉCIFIQUE DE SYSTÈMES D'INFORMATION DE GESTION

SOMMAIRE

Description du thème

Présentation de la ressource

Cette ressource permet aux élèves, placés dans le contexte d'une organisation, de découvrir les divers aspects de la gestion de contenu de masse.

Durées indicatives :

- Partie 1 : 60 min
- Partie 2 : 60 min
- Partie 3 : 30 min
- Partie 4 : 90 min

La ressource peut être utilisée de façon à découvrir les notions qui y sont abordées ou bien comme une première illustration dans la séquence pédagogique du professeur.

La ressource applique les préconisations du programme, à savoir l'articulation entre l'observation, l'analyse, la conceptualisation et l'interprétation au travers d'un cas d'organisation dont les besoins ont été simplifiés.

La mobilisation des outils et ressources d'environnements numériques sont indispensables dans ce thème où, particulièrement, les technologies sont au cœur des transformations.

Aspects didactiques

Si la plupart des questions peuvent être traitées individuellement par les élèves, certaines questions peuvent être traitées en binôme ou en groupe.

Cette ressource concerne le programme de l'enseignement de spécialité management, sciences de gestion et numérique, plus particulièrement celui de l'enseignement spécifique systèmes d'information de gestion. Sa mise en œuvre doit se faire en établissant des liens avec le tronc commun. Si deux professeurs se partagent le tronc commun et l'enseignement spécifique, il serait intéressant de travailler de façon collaborative les points du programme concernés.

Par exemple, une partie de cette ressource peut être travaillée dans le cadre de l'enseignement spécifique et exploitée également dans le cadre du tronc commun. Les élèves concernés peuvent exposer une synthèse de leurs travaux qui pourraient se poursuivre sous le regard des thèmes du tronc commun qui sont liés.

Repères dans les programmes de terminale

Programme de management, sciences de gestion et numérique : enseignement commun

Thème 1 : les organisations et l'activité de production de biens et de services		
1.1. Quels produits ou quels services pour quels besoins ?	Démarche *marketing*.	L'organisation possède et recherche des informations sur les consommateurs ou usagers et sur ses concurrents. Elle a recours à différents outils numériques de recueil et de traitement de l'information et peut se saisir des possibilités offertes par l'analyse des données massives (*big data*).
1.4. Les transformations numériques, une chance pour la production ?	Intégration des nouvelles technologies : informatique en nuage (*cloud computing*), objets connectés, intelligence artificielle, données ouvertes.	Aujourd'hui le développement des objets connectés permet de collecter des données (par le biais de capteurs dont des caméras) qui viennent enrichir les systèmes de production pour permettre une amélioration continue des processus. Des algorithmes permettent l'exploitation de données. L'apprentissage automatique (intelligence artificielle) permet d'améliorer leurs performances jusqu'à résoudre des questions ou accomplir des tâches pour lesquelles ils n'ont pas été conçus *a priori*. Cela permet de prévoir ou de simuler le comportement d'un équipement (utile pour sa maintenance par exemple), voire un comportement humain en donnant une possibilité de personnalisation de l'offre de services ou de produits.

Programme de management, sciences de gestion et numérique : enseignement spécifique de systèmes d'information de gestion

Thème 1 : organisation et numérisation		
3.1. Comment peut-on produire de l'information à partir de données ?	Nouvelles bases de stockage (données massives, lacs de données).	Les organisations doivent de plus en plus gérer des données aux formats multiples. Pour cela, des lacs de données (*data lakes*) permettent d'agréger des données de sources et de formats différents. Les principes des données massives (*big data*) peuvent ensuite être mis en *œuvre* (analyses descriptives, diagnostiques, prédictives et prescriptives) ainsi que le forage de données (*data mining*) afin de générer des données créatrices de valeur.
4.1. La standardisation facilite-t-elle la circulation des informations ?	Document : numérisation, structuration, indexation. Langage de définition de documents. Structuration de contenu documentaire : hyperlien, métadonnées, syndication, référencement. Gestion de contenu documentaire : fonctions, outils, moteur de recherche.	Une part importante de la vie d'une organisation et de ses échanges avec son environnement se traduit par la production et la circulation d'informations sous différentes formes (documents, courriels, dossiers, pages Web, etc.). Afin que ces données soient correctement gérées, il convient d'adopter une structuration précise et des langages standardisés (notamment des langages de balisage), en distinguant le contenu de sa présentation. Pour valoriser leur patrimoine informationnel, les organisations structurent la gestion de leur contenu, en y associant notamment des métadonnées qui favorisent son indexation.

Énoncé et contexte

La Bibliothèque nationale de France (BnF) est un établissement public ayant pour mission de collecter, conserver, enrichir et communiquer le patrimoine documentaire national (livres, documents, podcasts, applications, sites pédagogiques, médias…).

Chaque semaine, plusieurs milliers de documents sont numérisés par les équipes internes de la BnF et ses prestataires de services. Il est donc nécessaire de gérer de manière efficace cette masse considérable d'informations. Des méthodes (fonctions) et outils permettent de faciliter cette gestion de contenus (gestion de l'information).

Partie 1 - Collecter de l'information

La BnF a plusieurs sources d'enrichissement dont la principale est le dépôt légal. La BnF accroît aussi ses collections par des acquisitions auxquelles elle consacre une part importante de son budget :

- acquisitions courantes, notamment pour constituer une collection de référence dans le domaine étranger ;
- acquisitions prestigieuses, patrimoniales, pour lesquelles elle est parfois aidée par des mécènes.

Cette collecte nécessite la numérisation de nombreux ouvrages.

Proposition de questionnent

À partir de recherches effectuées sur des sites de confiance :

1. Indiquer en quoi consiste le dépôt légal et préciser les documents concernés par le dépôt légal.

À partir du **document 1** et de la vidéo intitulée BNF / numérisation de masse[1] :

2. Rappeler en quoi consiste la numérisation d'un document et indiquer qui réalise cette opération à la BnF.
3. Repérer les étapes qui permettent de passer d'un livre physique à un livre numérique et préciser les conséquences de cette numérisation.
4. Définir une métadonnée. Citer des exemples de métadonnées dans ce contexte.
5. Préciser les intérêts (autres que financiers) que peut avoir la numérisation d'une œuvre.

Concernant la sous-traitance d'une partie des numérisations :

6. Repérer les garanties importantes que la BnF peut exiger de ses prestataires.
7. Préciser la nature de(s) document(s) dans le(s)quel(s) ce type de garanties est spécifié.
8. Expliquer en quoi consiste l'opération de conversion en mode texte et son avantage pour l'utilisateur.
9. Préciser les cas pour lesquels la reconnaissance est rendue difficile.
10. Indiquer l'autre élément qui détermine la qualité de la numérisation.
11. Préciser le langage utilisé par le format Alto.

À partir du **document 2** :

12. Repérer les technologies qui permettent aujourd'hui une meilleure reconnaissance de l'écriture.
13. Expliquer en quoi le big data a permis d'améliorer la reconnaissance de l'écriture.

1. https://www.youtube.com/watch?v=bKmBm7Ry-GM

Partie 2 – Conserver l'information

Au fil des siècles, la BnF a développé des techniques appropriées à sa mission de conservation – qu'elle soit curative ou préventive (surveillance de l'état et protection des collections, conditions climatiques des magasins, restauration). Elle dispose pour cela de plusieurs ateliers spécialisés selon les types de documents et les techniques de conservation, ainsi que d'un laboratoire. Elle a également mis en place un système de préservation de ses données numériques.

Proposition de questionnement

À partir du **document 3** et de la page Prestation d'archivage numérique[2] du site de la BnF :

1. Expliquer pourquoi l'archivage est crucial. Repérer les moyens d'y parvenir et citer le nom du système d'archivage de la BnF.
2. Indiquer la façon de procéder pour repérer l'obsolescence d'un format et préciser l'action à réaliser qui en découle pour éviter cette obsolescence.
3. En plus de la pérennisation de la conservation des documents, indiquer les autres garanties du dispositif d'archivage de la BnF et leur intérêt.
4. Expliquer en quoi ces garanties permettent de lutter contre les « fake news ».

À partir des trois affiches suivantes :

- Info + intox = infox : la fausse nouvelle aujourd'hui
- Pour y voir plus clair : outils collectifs
- Pour y voir plus clair : outils individuels

5. Réaliser une présentation pour mettre en évidence : une fausse nouvelle d'actualité, des outils collectifs et individuels permettant de lutter contre ce fléau.
6. Préciser les types de documents stockés par la BnF et la conséquence de ce stockage.
7. Expliquer l'expression « Po (1 000 To) ».
8. Expliquer en quoi consiste l'ouverture du système d'archivage de la BnF et l'intérêt qu'elle représente pour la BnF.

Partie 3 – Diffuser l'information

La BnF assure l'accès à ses collections et offre un cadre de travail de qualité, sur place et en ligne. Elle est ouverte 71 heures par semaine et reçoit ses publics du lundi au dimanche sur cinq sites.

La BnF déploie également une offre en ligne importante qui répond, comme dans les espaces physiques, à des besoins et à des publics divers. Grâce à Gallica[3], sa bibliothèque numérique, la BnF permet l'accès gratuit à plus de 5 millions de documents.

Proposition de questionnement

À partir du **document 4** et de la vidéo intitulée Gallica en vidéo[4] :

1. Relever comment les documents numériques sont rendus accessibles.

La BnF a choisi la diffusion des ressources avec les formats suivants : JPEG2000, PDF, HTML, MP3, ePUB.

2. Rechercher les raisons de ce choix en mettant en avant les particularités de ces formats. Présenter la réponse sous forme de tableau.
3. Expliquer le principe de la « marque blanche ». Préciser l'intérêt de cette marque pour les partenaires de la BnF.

2. https://www.bnf.fr/fr/prestation-archivage-numerique
3. https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop
4. https://www.bnf.fr/fr/gallica-la-bibliotheque-numerique-de-la-bnf-et-de-ses-partenaires#bnf--gallica-en-vid-o

Partie 4 – Vers l'ouverture des données – L'usage des données massives – Les lacs de données

La profusion et la grande diversité des données stockées à la Bibliothèque nationale de France a poussé celle-ci à regrouper sur une même page toutes les informations issues de ses différents catalogues, ainsi que de sa bibliothèque numérique Gallica.

Le projet nommé data.bnf.fr[5] utilise les outils du Web sémantique[6] et s'inscrit dans une démarche d'ouverture des données[7]. Mis en ligne en juillet 2011, data.bnf.fr continue d'évoluer et de s'accroître.

Proposition de questionnement

À partir du **document 5** et de recherches sur Internet :

4. Définir une donnée publique.

À partir du site anthedesign.fr[8] et d'éventuelles recherches complémentaires :

5. Définir et expliquer la notion de web sémantique.

Sur le portail Data.bnf.fr, procéder à une recherche (de préférence sous FireFox) afin d'observer les métadonnées associées aux ressources concernant « Harry Potter » :

- Saisir « Harry Potter » dans la barre de recherche.
- Cliquer sur le premier lien obtenu dans la rubrique « Œuvres ».
- Cliquer en bas de page sur télécharger en JSON.

6. Expliquer le format de données JSON et son utilité.
7. Citer trois métadonnées obtenues. Préciser l'intérêt des métadonnées obtenues.

À partir de l'utilisation de la carte mondiale dynamique du portail Data.bnf.fr :

8. Retrouver le nombre de ressources associées à la ville de Bordeaux recensées sur le portail.
9. Expliquer en quoi cet outil illustre bien la notion de big data.

À partir de recherches sur Internet :

10. Citer d'autres exemples concrets mis en place par des entreprises pour gérer les données du Big Data.

À partir du **document 6** et de recherches sur Internet :

11. Expliquer pourquoi l'Ina a souhaité constituer un lac de données.

À partir des articles suivants, en annexes de cette ressource :

- Data Lake : définition, avantages et inconvénients pour l'entreprise
- SeLoger.com arrive à faire rimer RGPD avec Agilité

12. Citer des entreprises ayant mis en place un lac de données et préciser le but poursuivi.

5. https://data.bnf.fr/
6. https://data.bnf.fr/semanticweb
7. https://data.bnf.fr/licence
8. https://www.anthedesign.fr/autour-du-web/web-semantique/

Ressources

Document 1 : Du document papier au document numérique

Extraits du site de la Bibliothèque nationale de France[9].
La BnF numérise plus d'un million de pages par mois à partir de ses collections patrimoniales afin de faciliter l'accès à la culture.

Les ouvrages sont sélectionnés, indexés et traités :

- numérisation automatique pour les ouvrages brochés et massicotés ;
- numérisation manuelle pour les ouvrages fragiles ;
- transformation du « format image » – simple photographie – en « mode texte » – qui permet de réaliser des « copier-coller » et des recherches sémantiques dans le document – par le procédé d'océrisation (reconnaissance optique de caractères) ;
- vérification du résultat par le « contrôle qualité » ;
- mise en ligne (indexation) des ouvrages sur Gallica[10] et dans le catalogue de la BnF. L'index de Gallica est donc constitué à partir des métadonnées (titre, date ouvrage, auteur…), du contenu (plein texte) disponible, des tables des matières existantes, des légendes des images ;
- archivage dans Spar (le système de préservation d'archivage réparti) mis au point par la BnF.

La numérisation
Actuellement, un document numérisé est constitué des éléments suivants :

- des images au format JPEG 2000 en couleur ou en niveau de gris en résolution minimale à 400 DPI. Gallica permet de zoomer dans les images les plus grandes ;
- un manifeste : véritable fiche d'identité du document, il indique la pagination, l'historique des opérations de numérisation à fin de conservation, les légendes des images, etc. ;
- la table des matières avec les index, saisie en haute qualité afin de mieux parcourir le document dans Gallica et d'améliorer la recherche plein texte ;
- la reconnaissance optique de caractères qui permet la recherche plein texte. Lors de cette opération, la position du mot dans la page est repérée afin de permettre la surbrillance des occurrences recherchées dans Gallica. Le repérage des mots est compris dans les opérations de segmentation qui visent à établir la structure de l'ensemble du texte (mot, ligne de texte, bloc de texte, etc.).

Les prestataires de la numérisation – la qualité de la prestation
La BnF a mis en place, dès le lancement de la numérisation pour la constitution de la bibliothèque numérique Gallica, des outils et des procédures pour évaluer l'exhaustivité et la qualité des prestations exécutées dans le cadre des marchés de production d'images numériques. Elle les a ensuite améliorés et fait évoluer au cours du temps, en particulier pour la numérisation en nombre, qui nécessite une gestion fiable et efficace de gros volumes, sans pénaliser pour autant la qualité des documents numériques produits.

9. www.bnf.fr
10. Bibliothèque numérique de la BnF et de ses partenaires, accessible gratuitement sur Internet.

Pour assurer tous ces travaux de numérisation, la BnF s'appuie sur ses ateliers internes à hauteur de 20 % et sur des prestataires choisis dans le cadre de marchés publics. Les ateliers internes numérisent les documents spécifiques qui ne pourraient être confiés à un prestataire (grande fragilité, préciosité, etc.).

Les relations avec les fournisseurs ont été formalisées de manière très précise par un échange d'informations nécessaires à la production des différents types de données qui vont être exploitées par la BnF, tant pour l'archivage à long terme des documents numériques que pour la production des éléments à mettre en ligne.

Afin de garantir la qualité des données produites dans le cadre de ses marchés de numérisation de masse, la BnF a demandé à ses prestataires de fournir un plan assurance qualité (PAQ) lui permettant de s'assurer qu'ils avaient acquis une bonne compréhension de ses attentes.

La conversion en mode texte

Afin de répondre aux usages des internautes, la BnF assure la conversion en mode texte des documents imprimés le permettant et préalablement numérisés en mode image. Cette conversion est assurée automatiquement par un logiciel et fait l'économie de la retranscription manuelle, beaucoup plus chère. Les mots et chaînes de caractères stockés dans un fichier texte peuvent être réutilisés pour une nouvelle mise en page, exploités dans une base de données, etc.

Les techniques d'OCR[11] (*optical character recognition*) sont en progrès constant pour répondre à la demande très forte, mais la qualité de reconnaissance dépend malgré tout d'un grand nombre de facteurs liés tant au document original (contraste, défaut d'impression, mise en page en colonnes, polices trop petites ou trop grandes, alphabets non latins…) qu'à la numérisation elle-même.

Afin d'exploiter les résultats de l'OCR, on utilise à la BnF un format basé sur XML et géré par un schéma (document décrivant la structure à respecter pour le fichier XML), le format ALTO[12] (*Analyzed Layout and Text Object*) qui permet la segmentation d'une page en différents éléments composés de sous-éléments.

Pour chaque document numérisé par la BnF, le taux de qualité calculé automatiquement par le logiciel est vérifié manuellement par le prestataire sur un échantillon de mots, conformément à la norme ISO 2859-1. Cette opération permet de confirmer le taux de qualité annoncé.

Pour une partie des documents numérisés, la BnF exige un taux de qualité supérieur à 99,9 %. Pour tous ces documents, quel que soit le taux de qualité après OCR, le prestataire doit garantir ce taux en employant tous les moyens de corrections nécessaires, y compris manuels.

11. Reconnaissance optique des caractères.
12. Mise en page et objet de texte analysés.

Document 2 : La reconnaissance d'écriture

Extraits du site Data Analytics Post[13], site d'information sur les data sciences, porté par le master MVA de l'ENS Paris-Saclay

Reconnaître et comprendre une écriture met en jeux toutes les composantes de l'intelligence artificielle (IA) : il faut visualiser une image et détecter le texte (ce qui suppose de disposer de méthodes de perception visuelle), suivre le tracé de l'écriture (via un planning et le suivi d'une séquence d'actions) puis reconnaître les caractères (grâce à des algorithmes de reconnaissance de formes) et enfin reconnaître les mots et les phrases (par le traitement automatique de la langue) pour aller jusqu'à les comprendre (via une modélisation sémantique). C'est sans doute pour cette raison que la reconnaissance d'écriture partage avec la reconnaissance de la parole et la traduction automatique le privilège d'être parmi les plus anciens problèmes d'IA…

Dans la plupart des cas simples, les performances des systèmes de reconnaissance d'écriture sont aujourd'hui comparables à celles de l'humain, voire les dépassent, du moins pour les tâches les plus proches de la perception (détection du texte, suivi, reconnaissance des caractères et des mots). Pourtant, cela reste un véritable terrain d'expérimentation. La reconnaissance d'écriture manuscrite est même souvent considérée comme la drosophile des chercheurs en IA.

À partir de la fin des années 2000, la reconnaissance d'écriture a peut-être été le premier domaine à être profondément transformé par le renouveau des réseaux de neurones[14].

Un réseau de neurones artificiels, ou *Artificial Neural Network* en anglais, est un système informatique matériel et/ou logiciel dont le fonctionnement est calqué sur celui des neurones du cerveau humain.

Il aura fallu attendre le début des années 2010, avec l'essor du *Big Data* et du traitement massivement parallèle, pour que les *Data Scientists* disposent des données et de la puissance de calcul nécessaires pour exécuter des réseaux de neurones complexes. En 2012, lors d'une compétition organisée par ImageNet, un Neural Network est parvenu pour la première fois à surpasser un humain dans la reconnaissance d'image.

Par le biais d'un algorithme, le réseau de neurones artificiels permet à l'ordinateur d'apprendre à partir de nouvelles données. L'ordinateur doté du réseau de neurones apprend à effectuer une tâche en analysant des exemples pour s'entraîner. Ces exemples ont préalablement été étiquetés afin que le réseau puisse savoir ce dont il s'agit.

13. https://dataanalyticspost.com/la-reconnaissance-decriture-manuscrite-de-nouvelles-applications-pour-un-des-plus-vieux-problemes-dia/
14. https://dataanalyticspost.com/Lexique/reseau-de-neurones/

Document 3 : L'enjeu de l'archivage des documents

Extraits du site de la Bibliothèque nationale de France

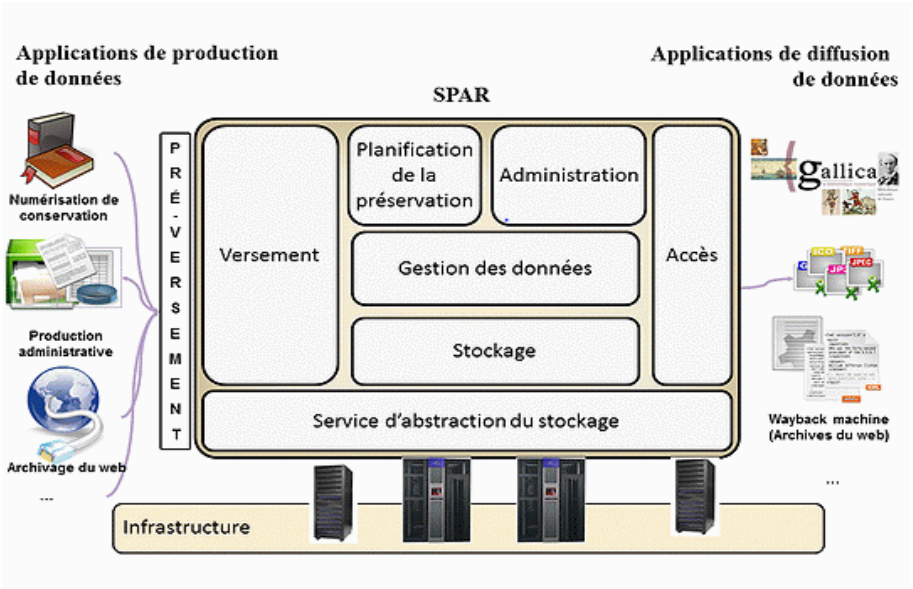
La conservation des données numériques par la BnF pose de façon cruciale la question de la pérennisation.

L'archivage à long terme ne se résume pas au stockage, mais nécessite la mise en place d'un dispositif plus complexe, capable de réaliser des opérations spécifiques, comme la migration de formats et de supports, qui assurent la lisibilité des documents à très long terme.

Spar est bien plus qu'un simple entrepôt de données sécurisé.

- Il permet de garantir la **continuité d'accès** en procédant aux transformations nécessaires en cas d'obsolescence technologique des outils informatiques de restitution. Ainsi, par exemple, lorsque le format d'image JPEG deviendra obsolète, Spar sera en mesure de transformer les images concernées dans un nouveau format plus performant.
  Actuellement, le format d'archivage des images est le TIFF[15] monopage non compressé et les formats de diffusion sont le PNG et le JPEG. Pour la presse, le format d'archivage est le TIFF monopage. Le format de diffusion est le JPEG2000. Ceci implique un travail permanent de veille technologique sur les formats et les outils.

Schéma présentant le fonctionnement du magasin numérique Spar © BnF



- Il garantit également, pour les données conservées :
  - **Intégrité :** cohérence du contenu :
    - Contrôle à l'entrée (empreinte des fichiers – checksum) ;
    - Audit régulier permettant de vérifier l'état des fichiers ;
    - Horodatage : historique des actions et des différentes versions ;

15. TIFF est un format, proche du BMP, offrant une image de très bonne qualité, mais qui est également très volumineuse. Développé par l'entreprise Microsoft, il appartient désormais à l'entreprise Adobe. Il est lisible par la plupart des logiciels de traitement d'images.

- **Authenticité :** gestion des droits et habilitations ;
- **Sécurité :**
  - Sécurité physique : conservation sur des serveurs redondés, situés sur deux sites distincts en France, sur des bandes et des disques, surveillance de l'état des équipements ;
  - Salles informatiques en accès restreint, plan de continuité et de reprise d'activité ;
  - Sécurité logique : étanchéité des serveurs, traçabilité des accès ;
  - Échanges et consultations des données archivées : possible à tout moment, via une interface dédiée.

Développé par la BnF pour ses collections numériques, le système d'archivage Spar (système de préservation et d'archivage réparti) permet de telles opérations sur un ensemble important de documents de tous types : y sont conservés des millions de documents numérisés (textes, images, musiques, vidéos), plusieurs milliards de pages web (issues du dépôt légal), représentant plusieurs Po (1 000 To) de données.

D'emblée s'est posée la question de l'ouverture de ce système à d'autres organisations, confrontées aux mêmes problèmes et souhaitant bénéficier des technologies et du savoir-faire de la BnF dans ce domaine. Ainsi, dans sa volonté de mutualiser les expertises et les coûts, la BnF propose aujourd'hui, à coûts maîtrisés, un service de tiers archivage comportant les mêmes garanties de sécurité et de pérennité que celles mises en œuvre pour ses propres collections patrimoniales.

La BnF dispose aujourd'hui d'un des systèmes hébergeant le plus gros volume de données en France : il est actuellement d'une capacité de plusieurs dizaines de Po. La présence des serveurs en France permet de donner des assurances fortes en matière de confidentialité des données et de respect des droits (souveraineté).

Document 4 : L'accès aux documents numériques – la diffusion

Extraits du site de la Bibliothèque nationale de France

Cette étape consiste à rattacher le document ainsi numérisé au catalogue de la BnF (consultable par l'internaute via le moteur de recherche).

Les formats offerts sur Gallica[16] ont été choisis par la BnF pour être les plus accessibles possible par le plus grand nombre en privilégiant les standards ouverts ou les standards de fait. Exemples : JPEG (2000), HTML, PDF, MP3, ePub, PDF…

Gallica est l'une des plus importantes bibliothèques numériques accessibles gratuitement sur l'internet. Elle offre l'accès à tous types de documents : imprimés (livres, presse et revues) en mode image et en mode texte, manuscrits, documents sonores, documents iconographiques, cartes et plans, vidéos.

Gallica s'adresse à tout lecteur, du curieux au bibliophile, du lycéen à l'universitaire.

Au 1er janvier 2020, Gallica proposait la consultation en ligne de 6 573 228 documents, dont 702 538 livres, 3 591 983 fascicules de presse et revues, 1 410 638 images, 134 087 manuscrits, 173 039 cartes, 50 291 partitions, 51 150 enregistrements sonores, 457 839 objets et 1 663 vidéos. Un certain nombre d'ouvrages a fait l'objet d'une reconnaissance optique de caractères[17] et le texte peut être recherché sur Gallica.

16. https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop
17. ROC, en anglais optical character recognition (OCR), ou océrisation, désigne les procédés informatiques pour la traduction d'images de textes imprimés ou dactylographiés en fichiers de texte.

À partir de 2013, dans le cadre des accords conclus par BnF-Partenariats, la BnF propose aux bibliothèques souhaitant diffuser leurs contenus sans disposer de leur propre outil, d'utiliser Gallica en « marque blanche ».

2018 a vu les développements de nouvelles fonctionnalités pour répondre aux attentes des usagers de Gallica et atteindre de nouveaux publics. Par exemple :

- les sites de la galaxie Gallica ont basculé vers le protocole HTTPS ;
- Gallica est désormais disponible en trois langues (français, anglais et italien) ;
- le moteur de recherche Exalead de Gallica est passé en version V6 plus, qui apporte une amélioration pour la recherche, notamment dans les titres.

Document 5 : *Big data* – le portail Data.bnf.fr

Extraits des sites Archimag[18] et de la Bibliothèque nationale de France

Doit-on l'appeler *big data* ? Data déluge ? Données massives ? Une chose est sûre, un véritable tsunami numérique s'est abattu sur nos entreprises et nos organisations. Chaque jour, nous produisons collectivement 2,5 trillions de données soit 1 milliard de milliard de données (10 puissance 18).

Selon une étude commandée par IBM, il apparaît que 90 % des données disponibles aujourd'hui ont été créées au cours des deux dernières années seulement !

À ce jour, peu d'institutions culturelles sont passées à l'exploitation opérationnelle de leurs données. Parmi celles qui ont franchi le pas, la Bibliothèque nationale de France fait figure de précurseur avec son portail Data.bnf.fr. Mise en ligne dès le mois de juillet 2011, cette plateforme a pour ambition d'accroître la visibilité sur le web des innombrables ressources documentaires détenues par la BnF : catalogues, notices, documents numérisés…

Une initiative bienvenue car peu d'internautes connaissent l'existence de ce patrimoine numérique à « forte valeur ajoutée ».

Le portail data.bnf.fr permet :

- d'accéder aux ressources de la BnF directement depuis une page Web, sans avoir à connaître préalablement les services de la BnF ;
- de s'orienter dans les ressources de la BnF et de trouver éventuellement des ressources extérieures.

L'objectif est donc de valoriser la richesse des fonds de la BnF sur le Web et de servir de pivot entre les différentes ressources : data.bnf.fr est donc au service des autres applications de la BnF. Enfin, le projet s'inscrit dans une démarche d'ouverture de la BnF au Web de données et d'adoption des standards du Web sémantique[19].

Le projet data.bnf.fr se place ainsi résolument dans le mouvement d'ouverture des données publiques (*Open Data*). Portée par des acteurs civiques et les gouvernements, l'ouverture des données publiques vise à rendre accessibles les données non nominatives, ne relevant ni de la vie privée, ni de la sécurité et collectées ou produites par des organismes publics.

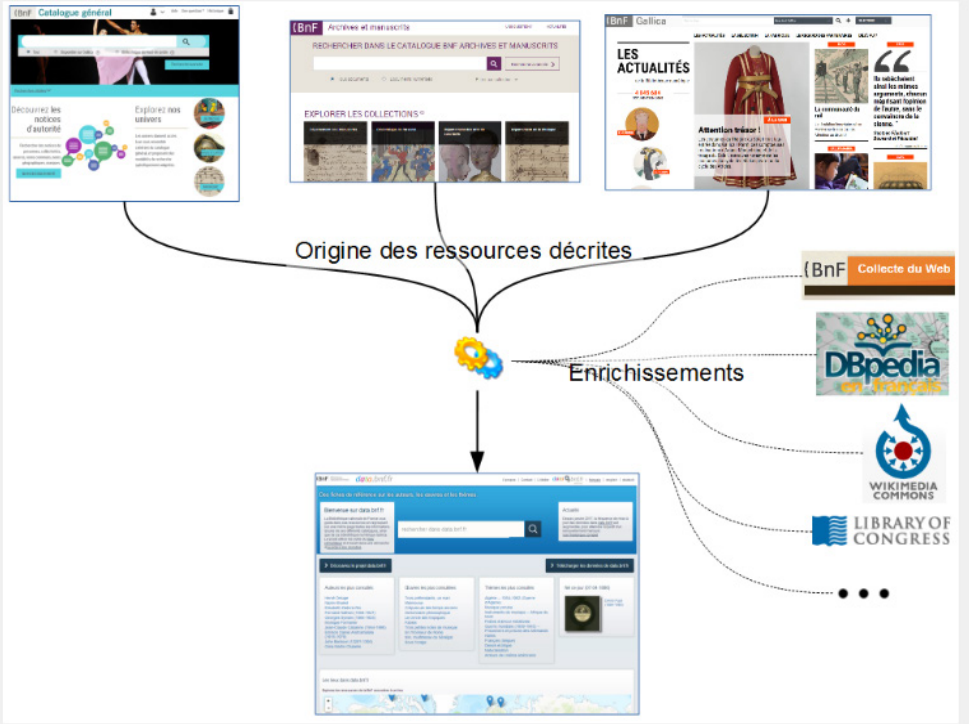
18. https://www.archimag.com/chiffre-du-jour/2015/11/26/big-data-aller-ou-pas
19. https://data.bnf.fr/semanticweb

Comment ça marche ?

Data.bnf.fr extrait, transforme et regroupe dans une base commune des données issues de bases distinctes et produites dans des formats différents, afin de les lier entre elles et de les rendre interopérables.

Ses pages sont indexées par les moteurs de recherche, alors que ceux-ci ne référencent pas les données et les métadonnées qui sont cachées dans les bases non indexables de la BnF. Les pages de data.bnf.fr décrivent les ressources de la BnF qui sont souvent dissimulées dans le Web « profond » et signalent les documents numériques directement accessibles.

Les liens externes dans data.bnf.fr



Après plusieurs années d'exploitation, la valorisation des données semble donner les résultats escomptés. Au mois de novembre 2014, le portail Data.bnf.fr recouvrait plus de 60 % des catalogues de la BnF, soit environ 7 millions de documents issus du catalogue général et de l'entité « archives et manuscrits » de l'établissement.

À terme, la plateforme devrait intégrer un impressionnant volume de données de qualité : plus de 15 millions de données d'autorité et bibliographiques. « Cet accroissement du volume du site implique des évolutions techniques (performance, mise à jour des données) et ergonomiques du site », explique la Bibliothèque nationale de France. Un enjeu technique et documentaire d'autant plus important que la BnF doit aligner ses référentiels sur d'autres jeux de données du web, notamment ceux produits par d'autres institutions publiques françaises. Et, à terme, offrir à ce patrimoine numérique l'audience qu'il mérite.

Document 6 : *Big data* – Lac de données

Extraits du site Didak'TIC[20]

Actuellement, l'Ina (Institut national de l'audiovisuel) s'engage dans une démarche de stockage des données orientées Big data. Il s'agit de fondre tous les systèmes documentaires au sein d'un lac de données dans le but de rassembler toutes les données que l'institution conserve (ce qui inclut les « métadonnées documentaires, commerciales, juridiques et d'usage »)…

Un lac de données (*Data Lake* en anglais) est défini par l'absorption de flux de données brutes rendues utilisables pour analyse. Des données disparates sont collectées puis stockées en continu dans un espace que l'on pourrait qualifier de « réservoir ».

Schématiquement, une base de données relationnelle est une structure verticale difficile à déconstruire si l'on souhaite en modifier l'organisation.

Un peu comme un gratte-ciel, si votre entrepôt de données prend de la hauteur et conserve de plus en plus de données, sa déconstruction devient problématique si vous souhaitez changer d'angle d'analyse. Un lac de données est à l'inverse totalement plat, sans structure. Les données sont conservées sur le même plan. La structure est alors créée au moment de l'analyse. On parle de « data lake », mais aussi de « data réservoir », réservoir de données.

20. https://www.didaktic.fr/big-data/archivage_web/2/

Guide d'accompagnement pédagogique

Il s'agit ici d'éléments de réponse aux propositions de questionnement.

Partie 1 - Collecter de l'information

À partir de recherches effectuées sur des sites de confiance :

1. Indiquer en quoi consiste le dépôt légal et préciser les documents concernés par le dépôt légal.

https://www.bnf.fr/fr/quest-ce-que-le-depot-legal

En France, le dépôt légal est l'obligation pour tout éditeur, imprimeur, producteur, importateur, de déposer chaque document qu'il édite, imprime, produit ou importe, auprès de l'organisme habilité à recevoir le dépôt en fonction de la nature du document. Cette obligation s'applique à tout document diffusé en nombre à un public s'étendant au-delà du cercle de famille.

https://www.bnf.fr/fr/depot-legal-pour-quels-documents

Sont soumis au dépôt légal tous les documents mis en nombre à la disposition d'un public, à titre onéreux ou gratuit. Cette obligation s'applique aux documents imprimés, mais aussi aux cartes et plans, photographies, partitions musicales, documents audiovisuels, logiciels, bases de données, jeux vidéo, etc.

À partir du **document 1** et de la vidéo intitulée BNF / numérisation de masse[21] :

2. Rappeler en quoi consiste la numérisation d'un document et indiquer qui réalise cette opération à la BnF.

La numérisation est une opération qui consiste à récupérer sur support numérique (ou dématérialiser) un document initialement possédé sur support papier. Cette numérisation est effectuée en partie (20 %) en interne. La numérisation du reste des documents est confiée à des prestataires de services. Les ateliers internes numérisent les documents fragiles qui ne peuvent être confiés à un prestataire. D'autres bibliothèques partenaires de la BnF numérisent 30 % des images dans le cadre du principal marché de numérisation des livres reliés.

3. Repérer les étapes qui permettent de passer d'un livre physique à un livre numérique et préciser les conséquences de cette numérisation.

1. Numérisation automatique pour les ouvrages brochés et massicotés et manuelle pour les ouvrages fragiles ;
2. transformation du « format image » en « mode texte », ce qui permet de « copier-coller » et d'effectuer des recherches sémantiques dans le document ;
3. vérification du résultat par le « contrôle qualité » ;
4. mise en ligne des ouvrages sur Gallica et dans le catalogue de la BnF ; indexation avec des métadonnées ;
5. archivage.

21. https://www.youtube.com/watch?v=bKmBm7Ry-GM

Ces étapes entrainent un coût et des temps de traitement importants et nécessitent des espaces de stockage conséquents.

4. Définir une métadonnée. Citer des exemples de métadonnées dans ce contexte.
Une métadonnée est une donnée servant à définir ou décrire une autre donnée. On peut citer, comme exemple, les informations permettant de caractériser un document : auteur, nombre de pages, catégorie du document…

5. Préciser les intérêts (autres que financiers) que peut avoir la numérisation d'une œuvre.

- Faire connaître les ouvrages anciens et/ou méconnus par une plus grande part de la société (tout profil confondu) ;
- Assurer sa sauvegarde et sa pérennité ;
- Participer à la mémoire nationale ;
- Effectuer des recherches plein texte dans le document ;
- Assurer son enrichissement (via des métadonnées).

Concernant la sous-traitance d'une partie des numérisations :

6. Repérer les garanties importantes que la BnF peut exiger de ses prestataires.

- Confidentialité des contenus numérisés ;
- Qualité de la numérisation et de la description (respect du schéma XML) ;
- Respect des délais.

7. Préciser la nature de(s) document(s) dans le(s)quel(s) ce type de garanties est spécifié.
Un plan assurance qualité (PAQ) doit être fourni par le prestataire afin d'attester qu'il a acquis une bonne compréhension des attentes de la BnF. De plus, un contrat de sous-traitance doit être signé entre les parties.

8. Expliquer en quoi consiste l'opération de conversion en mode texte et son avantage pour l'utilisateur.
La conversion en mode texte consiste à repérer et reconnaître des mots pour faire une recherche plein texte ou un copier-coller pour une utilisation dans un autre logiciel.

9. Préciser les cas pour lesquels la reconnaissance d'écriture est rendue difficile.
Un contraste insuffisant, un défaut d'impression, une mise en page en colonnes, des polices trop petites ou trop grandes, des alphabets non latins rendent difficile la reconnaissance d'écriture.

10. Indiquer l'autre élément qui détermine la qualité de la numérisation.
La qualité (efficacité) du matériel de numérisation (résolution, qualité de l'OCR) est déterminante.

11. Préciser le langage utilisé par le format Alto.
Il s'agit du langage XML dont un schéma[22] est fourni pour décrire la structure à respecter.

> **Pour aller plus loin**
> Numérisation de masse : qualité et formats utilisés pour garantir la conservation
> https://www.bnf.fr/fr/numerisation-de-masse-qualite-et-formats-utilises-pour-garantir-la-conservation

22. https://www.bnf.fr/fr/techniques-et-formats-de-conversion-en-mode-texte#bnf-le-format-alto

À partir du **document 2** :

12. Repérer les technologies qui permettent aujourd'hui une meilleure reconnaissance de l'écriture.

Reconnaître et comprendre une écriture met en jeux toutes les composantes de l'intelligence artificielle (IA) : il faut visualiser une image et détecter le texte (ce qui suppose de disposer de méthodes de perception visuelle), suivre le tracé de l'écriture (via un planning et le suivi d'une séquence d'actions) puis reconnaître les caractères (grâce à des algorithmes de reconnaissance de formes) et enfin reconnaître les mots et les phrases (par le traitement automatique de la langue) pour aller jusqu'à les comprendre (via une modélisation sémantique). C'est sans doute pour cette raison que la reconnaissance d'écriture partage avec la reconnaissance de la parole et la traduction automatique le privilège d'être parmi les plus anciens problèmes d'IA.

13. Expliquer en quoi le *big data* a permis d'améliorer la reconnaissance de l'écriture.

L'essor du big data a permis aux scientifiques de disposer des données et de la puissance de calcul nécessaires pour exécuter des réseaux de neurones complexes. Un réseau de neurones artificiels est un système informatique matériel et/ou logiciel dont le fonctionnement est calqué sur celui des neurones du cerveau humain. Par le biais d'un algorithme, le réseau de neurones artificiels permet à l'ordinateur d'apprendre à partir de nouvelles données. L'ordinateur doté du réseau de neurones apprend à effectuer une tâche en analysant des exemples pour s'entraîner.

Partie 2 – Conserver l'information

À partir du document 3 et de la page Prestation d'archivage numérique[23] du site de la BnF :

1. Expliquer pourquoi l'archivage est crucial. Repérer les moyens d'y parvenir et citer le nom du système d'archivage de la BnF.

L'enjeu de l'archivage est la pérennisation de consultation du document stocké à long terme. Cela nécessite la migration de formats et l'usage de supports pérennes. Le système élaboré par la BnF est appelé Spar.

2. Indiquer la façon de procéder pour repérer l'obsolescence d'un format et préciser l'action à réaliser qui en découle pour éviter cette obsolescence.

Pour se rendre compte qu'un format est obsolète, il faut procéder à la veille technologique sur les formats de stockage. Ainsi, on est au courant de l'utilisation de nouveaux formats de stockage et de l'abandon d'anciens formats. Dans ce cas, il est alors nécessaire de convertir tous les documents numérisés dans le nouveau format pour assurer la lisibilité à long terme.

Exemples :

- Les formats doc, xls, ptt remplacés par les formats (Office) Open XML (docx, xslx, pptx)
- Vers la fin du jpeg : https://arts.konbini.com/partners/vers-la-fin-du-format-jpeg
- Le format mp3 devient officiellement obsolète : https://www.rtbf.be/tendance/techno/detail_le-format-mp3-devient-officiellement-obsolete?id=9606302

23. https://www.bnf.fr/fr/prestation-archivage-numerique

3. En plus de la pérennisation de la conservation des documents, indiquer les autres garanties du dispositif d'archivage de la BnF et leur intérêt.
Autres garanties apportées par Spar : intégrité, authenticité et sécurité des données. Ce sont là trois règles nécessaires pour que le document ait une valeur juridique : celle-ci découle du fait que le document numérisé est maintenu dans son intégrité. Dès lors que l'intégrité d'un document est assurée, il peut servir aux mêmes fins et produire les mêmes effets juridiques que le document sur support-papier dans les situations où il respecte les règles de droit qui lui sont applicables.

4. Expliquer en quoi ces garanties permettent de lutter contre les « fake news ».
Une fausse nouvelle, que l'actualité remet au goût du jour sous l'appellation de *fake news* ou d'infox (« information » et « intoxication »), est un phénomène très ancien contre lequel doit lutter au quotidien la BnF. Garantir l'intégrité d'une œuvre permet sans nul doute d'éviter certaines « fausses nouvelles » à propos de cette œuvre, mais le fléau des *fake news* est loin de pouvoir être résolu à ce jour.

> **Pour aller plus loin :**
> Afin de sensibiliser le public aux *fake news*, la BnF a organisé en 2019 un colloque durant la Semaine de la presse à l'école, lancé de nouveaux ateliers pédagogiques et proposé une exposition sur panneaux (https://www.bnf.fr/fr/agenda/les-democraties-lepreuve-des-infox).

À partir des trois affiches suivantes :

- Info + intox = infox : la fausse nouvelle aujourd'hui
- Pour y voir plus clair : outils collectifs
- Pour y voir plus clair : outils individuels

5. Réaliser une présentation pour mettre en évidence : une fausse nouvelle d'actualité, des outils collectifs et individuels permettant de lutter contre ce fléau.
Ce travail peut être réalisé en groupe.

6. Préciser les types de documents stockés par la BnF et la conséquence de ce stockage.
Il s'agit de documents numérisés (textes, images, musiques, vidéos) et de pages web (issues du dépôt légal). Ce stockage demande un espace de stockage très important.

7. Expliquer l'expression « Po (1 000 To) ».
Il s'agit de la capacité de disque nécessaire au stockage des documents de la BnF. L'unité de mesure de base est l'octet, mais, étant donné les grandeurs manipulées au fil du temps, des multiples ont été créés : Ko, Mo, Go, To (Téraoctet), Po (Pétaoctet) avec un coefficient de multiplication de 1 000 à chaque passage d'un multiple à l'autre.

8. Expliquer en quoi consiste l'ouverture du système d'archivage de la BnF et l'intérêt qu'elle représente pour la BnF.
La BnF propose un service qui permet à d'autres organisations, confrontées aux mêmes problèmes d'archivage et souhaitant bénéficier des technologies et du savoir-faire de la BnF dans ce domaine, d'accéder à un système d'archivage comportant les mêmes garanties de sécurité et de pérennité que celles mises en œuvre pour ses propres collections patrimoniales. Cette ouverture permet une mutualisation des expertises et des coûts afin d'amortir plus rapidement l'investissement de la BnF pour la réalisation de ce système et ainsi permettre son évolution prochaine (compte tenu de l'évolution rapide des technologies).

Partie 3 – Diffuser l'information

À partir du document 4 et de la vidéo intitulée Gallica en vidéo[24] :

1. Relever comment les documents numériques sont rendus accessibles.

Les documents numériques sont rattachés au catalogue pour être référencés dans le moteur de recherche de la BnF et accessibles à l'internaute via Gallica.

La BnF a choisi la diffusion des ressources avec les formats suivants : JPEG2000, PDF, HTML, MP3, ePUB.

2. Rechercher les raisons de ce choix en mettant en avant les particularités de ces formats. Présenter la réponse sous forme de tableau.

Format	Définition - particularités	Raisons de ce choix
JPEG2000	Format d'image qui s'intègre facilement dans une page web car il a un bon rapport qualité/poids grâce à la compression. Ses performances en compression sont supérieures à celle de JPEG (acronyme de Joint Photographic Experts Group).	On obtient des fichiers d'un poids inférieur à qualité d'image égale.
PDF	Format d'échange (consultation d'écran, impression, etc.) et d'archivage de documents électroniques.	Ce format est devenu un « standard international ».
HTML	HyperText Markup Language – page web statique.	Ce format permet de créer des documents interopérables[25] avec des équipements très variés de manière conforme aux exigences de l'accessibilité du web.
MP3	Format de compression audio avec perte.	Cette perte permet une réduction importante de la taille du flux de données audio, tout en conservant une qualité de restitution couramment jugée acceptable.
ePub	electronic publication - Protocole de communication, d'interconnexion ou d'échange et tout format de données interopérable dont les spécifications techniques sont publiques et sans restriction d'accès ni de mise en œuvre.	C'est un format ouvert standardisé pour les livres numériques. Il est fondé sur le XML.

Tous ces formats permettent un accès aisé, rapide et universel sans installation de logiciels supplémentaires de la part des internautes.

24. https://www.bnf.fr/fr/gallica-la-bibliotheque-numerique-de-la-bnf-et-de-ses-partenaires#bnf--gallica-en-vid-o-
25. Capacité que possède un système informatique à fonctionner avec d'autres produits ou systèmes informatiques, existants ou futurs, sans restriction d'accès ou de mise en œuvre.

3. Expliquer le principe de la « marque blanche ». Préciser l'intérêt de cette marque pour les partenaires de la BnF.

Une marque blanche est un service ou un produit conçu par une entreprise que d'autres entreprises reprennent à leur compte et commercialisent sous leur propre marque. Il s'agit donc d'un mécanisme commercial de mise à disposition d'outils ou de produits, sans citer la marque ni l'origine de l'information transmise. La particularité du dispositif Gallica « marque blanche » repose sur le fait que la bibliothèque numérique du partenaire est construite sur la base de l'infrastructure Gallica (le socle technique et les serveurs de diffusion sont communs). La bibliothèque du partenaire bénéficie ainsi de toutes les fonctionnalités actuelles et futures de Gallica tout en étant paramétrable et personnalisée aux couleurs du partenaire qui peut ainsi diffuser son contenu sans toutefois faire mention ni de la BnF ni de Gallica. Les pages sont paramétrables et utilisent la charte graphique spécifique du partenaire. Ainsi, malgré l'apparente filiation créée par la proximité ergonomique avec Gallica, l'utilisateur accède à un environnement spécifique qui porte le nom et l'identité graphique du partenaire.

Partie 4 – Vers l'ouverture des données – L'usage des données massives – Les lacs de données

À partir du document 5 et de recherches sur Internet :

1. Définir une donnée publique.

La notion de « donnée publique » couvre l'ensemble des informations ou données produites ou reçues par une autorité administrative dans le cadre de sa mission de service public. Elles sont communicables à toute personne en faisant la demande. Ces informations doivent être présentées sous un format permettant leur traitement automatisé et leur réutilisation.

À partir du site anthedesign.fr et d'éventuelles recherches complémentaires :

2. Définir et expliquer la notion de web sémantique.

Le Web sémantique est une extension du Web standardisée par le *World Wide Web Consortium* (W3C). Ces standards encouragent l'utilisation de formats de données et de protocoles d'échange normés sur le Web, en s'appuyant sur le modèle *Resource Description Framework* (RDF). Le Web sémantique est par certains qualifié de web 3.0. Selon le W3C, le Web sémantique fournit un modèle qui permet aux données d'être partagées et réutilisées entre plusieurs applications, entreprises et groupes d'utilisateurs (Wikipedia). Les sites Web deviennent des applications en ligne qui savent analyser automatiquement les contenus écrits et picturaux, qui savent les interpréter, les comprendre, les classer et les rediffuser vers un nouveau public internaute. L'idée est donc de permettre une recherche intelligente sur le Web, faite par des ordinateurs et basée sur des définitions qu'ils puissent « comprendre », des définitions données pour le monde entier. En faisant une requête sur un moteur proposant de la recherche en langage naturel, vous l'interrogerez comme vous parlez, et il transformera cette demande en langage compréhensible et cohérent pour la machine.

Sur le portail Data.bnf.fr, procéder à une recherche (de préférence sous FireFox) afin d'observer les métadonnées associées aux ressources concernant « Harry Potter » :

- Saisir « Harry Potter » dans la barre de recherche.
- Cliquer sur le premier lien obtenu dans la rubrique « Œuvres ».
- Cliquer en bas de page sur télécharger en JSON.

3. Expliquer le format de données JSON et son utilité.

Résultat indicatif de la recherche

https://data.bnf.fr/fr/16675365/harry_potter.json

JSON | Données brutes | En-têtes

Enregistrer | Copier | Tout réduire | Tout développer | Filtrer le JSON


```
▾ 0:
    modification-date:    "2020-05-20"
    url:                  "https://data.bnf.fr/fr/16675365/harry_potter/"
  ▾ title:                "Harry Potter - Œuvre - Ressources de la Bibliothèque nationale de France"
  ▾ description:          "Toutes les informations de la Bibliothèque Nationale de France sur : Harry Potter"
    site_name:            "data.bnf.fr"
    type:                 "book"
  ▾ image:                "https://data.bnf.fr/data/770f5349e9884a1771ae3805e26f0a0d/pasdevisuel2.jpg"
    publication:          "2001"
    lang:                 []
    notice-type:          "TIC"
    notes:                []
    ark:                  "ark:/12148/cb166753653"
    label:                "Harry Potter"
    manifestations:       []
    contributors:         []
```


JSON (JavaScript Object Notation – Notation Objet issue de JavaScript) est un format léger d'échange de données. Il est facile à lire ou à écrire pour des humains. Il est aisément analysable ou générable par des machines. Il est fondé sur un sous-ensemble du langage de programmation JavaScript. Il permet de représenter de l'information structurée (métadonnées) comme le permet XML par exemple (Wikipédia).

Pour aller plus loin

https://www.json.org/json-fr.html

4. Citer trois métadonnées obtenues. Préciser l'intérêt des métadonnées obtenues.
Métadonnées : url, titre, description.

On obtient des données structurées, exploitables par une application informatique, c'est-à-dire de l'open data.

À partir de l'utilisation de la carte mondiale dynamique du portail Data.bnf.fr :

5. Retrouver le nombre de ressources associées à la ville de Bordeaux recensées sur le portail.

- Lien consulté le 10/12/2020 :
- https://data.bnf.fr/fr/11977433/15241838/bordeaux__gironde__france_/
- Documents sur ce lieu (2 301 ressources dans data.bnf.fr).

6. Expliquer en quoi cet outil illustre bien la notion de big data.
On peut facilement voir la profusion des ressources proposées par cet outil. Les recherches peuvent prendre du temps dans un tel volume de données. De plus les données sont proposées dans des formats de plus en plus variés. Il s'agit de plus en plus de données non structurées. Ces formats variés rendent les traitements plus complexes, comme l'illustre bien le site de la BnF.

> **Faire un lien avec le Document 3**
> « La BnF dispose aujourd'hui d'un des systèmes hébergeant le plus gros volume de données en France : il est actuellement d'une capacité de plusieurs dizaines de Po. La présence des serveurs en France permet de donner des assurances fortes en matière de confidentialité des données et de respect des droits (souveraineté). »

À partir de recherches sur Internet :

7. Citer d'autres exemples concrets mis en place par des entreprises pour gérer les données du Big Data.
Source : Big Data et Data marketing[26], article publié le 14/10/2015 sur le site de l'académie de Versailles, Centre de Ressources en Économie-Gestion

Le constructeur automobile PSA s'est allié à IBM29 pour exploiter les données des véhicules connectés (1,5 millions de voitures). Grâce à l'analyse des usages, PSA peut améliorer la conception et la qualité des véhicules ainsi qu'offrir des services personnalisés aux clients.

La compagnie d'assurance Allianz propose à ses clients automobiles d'embarquer un boitier analysant les conditions et comportements de conduite. Celui-ci leur permet d'offrir des services d'assistance et conseils, mais également de bénéficier éventuellement d'une réduction de tarif pour bonne conduite.

SEB et Coheris se sont alliés pour imaginer une transformation numérique de la cuisine et créer un moteur de recommandation de recettes tenant compte de diverses données en provenance des objets connectés de la cuisine, mais aussi du profil, de la météo, de « like » de recettes issus de Facebook.

26. Article consulté le 10/12/2020 : https://creg.ac-versailles.fr/Big-Data-et-Data-marketing ou https://creg.ac-versailles.fr/IMG/pdf/big-data-et-data-marketing.pdf

Source : Les « big data », nouvel outil contre les épidémies comme Ebola ?[27], article publié le 27/10/2014 sur le site Sciences et Avenir

La société HealthMap, spécialisée en *big data* de santé, aurait détecté l'épidémie d'Ebola avant que l'OMS n'en fasse l'annonce officielle.

À partir du document 6 et de recherches sur Internet :

8. Expliquer pourquoi l'Ina a souhaité constituer un lac de données.

Lorsque le besoin est de stocker de gros volumes de données à structures variables, mais surtout dont on ne sait pas à l'avance comment elles vont être utilisées et analysées apparaît le concept de lac de données. Ainsi, l'Ina veut rassembler toutes ses données au sein d'un lac de données dans le but de créer un contenu dynamique, plus facilement exploitable.

À partir des articles suivants, en annexes de cette ressource :

- Data Lake : définition, avantages et inconvénients pour l'entreprise
- SeLoger.com arrive à faire rimer RGPD avec Agilité

9. Citer des entreprises ayant mis en place un lac de données et préciser le but poursuivi.

Source : *Data Lake* : définition, avantages et inconvénients pour l'entreprise

De plus en plus d'institutions scolaires s'intéressent aux *Data Lakes*. Par exemple, la Cardiff University a mis en place le projet *Personal Data Lake*, afin de créer un nouveau type de *Data Lake* permettant de gérer le *Big Data* d'utilisateurs individuels en leur fournissant un point centralisé de collecte, d'organisation et de partage de données personnelles.

Les entreprises qui ont recours à l'Internet des Objets sont très friandes du modèle *Data Lake*. En effet, il faut pouvoir rassembler les données en provenance de centaines, voire de millions de capteurs et les corréler. Cette infrastructure est par exemple au cœur du fonctionnement des compteurs connectés Linky en cours d'installation dans toute la France par Enedis. Ces derniers relèvent différents types d'informations sur la consommation, la puissance allouée, les défauts de sécurité et facilitent l'intervention des équipes de maintenance.

Source : SeLoger.com arrive à faire rimer RGPD avec Agilité

Le Groupe SeLoger est une entreprise de 800 personnes (dont 250 développeurs), qui gère une quinzaine de marques, avec des sites web et des applications mobiles pour chacune d'elles, et des données acquises via des opérations de croissance externe.

Alors que chaque site Web du groupe gérait ses propres données, priorité est maintenant donnée à une centralisation. Un *Data Lake* est créé en implémentant, dès sa conception, les règles du RGPD.

« Lorsque quelqu'un accède à la liste des inscrits pour une alerte immobilière, il y aura bien une ligne par inscrit, mais les noms, prénoms et emails seront chiffrés. La clé de chiffrement est en accès restreint et dès qu'un internaute est inactif depuis 3 ans, sa clé est détruite, donc automatiquement les milliers de lignes qui le concernent dans le Data Lake sont anonymisées, car nous ne sommes plus techniquement capables de retrouver le nom ».

27. Article consulté le 10/12/2020 : https://www.sciencesetavenir.fr/sante/les-big-data-nouvel-outil-contre-les-epidemies-comme-ebola_28006