Gestion de contenu et *big data*

Description du thème

Propriétés	Description
Intitulé long	Comprendre la gestion de contenu de masse.
Formation(s) concernée(s)	Terminale STMG.
Matière(s)	**Terminale STMG** • Management, sciences de gestion et numérique • Enseignement spécifique de SIG
Présentation	A travers la multitude de documents gérés et mis à la disposition par la Bibliothèque nationale de France, comprendre les enjeux de la gestion des données de masse.
Savoirs	Gestion de contenu et données massives.
Compétences	**Terminale STMG : management des sciences de gestion et numérique** Thème 1 : Les organisations et l'activité de production de biens et de services Q1.1 Quels produits et quels services pour quels besoins ? Q1.4 : Les transformations numériques, une chance pour la production ? **Terminale STMG : enseignement spécifique de SIG** Thème 3 : information, action et décision Q3.1 : Comment peut-on produire de l'information à partir de données ? Thème 4 : système d'information et échange Q4.1 : La standardisation facilite-t-elle la circulation des informations ? **Terminale STMG : enseignement spécifique de Mercatique** Thème 3 : La communication de l'offre Q3.2 Comment enrichir la relation client grâce au numérique ?
Transversalité	
Prérequis	**Première STMG : Sciences de gestion et numérique** Thème 2 : Numérique et intelligence collective Q2.1 : En quoi les technologies transforment-elles l'information en ressource ?
Outils	
Mots-clés	Numérisation, structuration, indexation, gestion de contenu documentaire, structuration de contenu documentaire, dématérialisation, métadonnée, données massives, *big data*, lacs de données, intelligence artificielle, *fake news* .
Durée indicative	4 heures – A traiter sur minimum 2 séances
Auteur(es)	Estelle CYBULA-SORNETTE
Re-lecteur(s)	Sébastien Henriot, Valéry Tschaen
Version	v 1.0

Introduction

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).

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.

À partir de recherches effectuées sur des sites de confiance :

1. **Indiquer en quoi consiste le dépôt légal ? Quels documents sont concernés ?**

À partir de l'annexe 1 et de la vidéo jointe (BNF - la numérisation de masse.mp4) – ou disponible à cette adresse *https://www.youtube.com/watch?v=bKmBm7Ry-GM* :

2. **Rappeler en quoi consiste la numérisation d'un document et indiquer qui réalise cette opération à la BnF.**
3. **Indiquer quelles étapes permettent de passer d'un livre physique à un livre numérique ? Quelles en sont les conséquences ?**
4. **Rappeler ce qu'est une métadonnée ? Citer des exemples de métadonnées dans ce contexte.**
5. **Dire quels intérêts (autres que financier) peuvent apporter la numérisation d'une œuvre ?**
6. **Concernant la sous-traitance d'une partie des numérisations, indiquer :**
   - 6.1. **Quelles peuvent être les garanties importantes à exiger des prestataires de la BNF ?**
   - 6.2. **Dans quel(s) document(s) spécifie-t-on ce type de garanties ?**
7. **Expliquer en quoi consiste l'opération de conversion en mode texte ? Quel est son avantage pour l'utilisateur ?**
8. **Dire dans quels cas la reconnaissance est rendue difficile ?**
9. **Indiquer quel autre élément détermine la qualité de la numérisation ?**
10. **Indiquer quel est le langage utilisé par le format Alto ?**

À partir de l'annexe 2 :

11. **Dire quelles technologies permettent aujourd'hui une meilleure reconnaissance de l'écriture ?**

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

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.

À partir de l'annexe 3 et du site de la BNF *https://www.bnf.fr/fr/prestation-archivage-numerique* **:**

13. **Expliquer pourquoi l'archivage est crucial ? Comment y parvenir ? Quel est le nom du système d'archivage de la BNF ?**

14. **Indiquer comment a-t-on connaissance de l'obsolescence d'un format ? Quelle est alors l'action réalisée ?**

15. **Indiquer quelles sont les autres garanties du dispositif Spar ? Dans quel but ?**

16. **Dire en quoi ces garanties permettent de lutter contre les « *fake news* » ?**

À partir des 3 affiches fournies :

17. **Réaliser une présentation permettant de mettre en évidence la définition d'une fausse nouvelle d'actualité accompagnée d'exemples et présenter des outils collectifs et individuels permettant de lutter contre ce fléau.**

18. **Indiquer quels sont les types de documents stockés par la BnF ? Quelle est la conséquence de ce stockage ?**

19. **Expliquer l'expression « Po (1 000 To) ».**

20. **Dire en quoi consiste l'ouverture du système Spar? Quel est l'intérêt pour la BNF ?**

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, sa bibliothèque numérique, la BnF permet l'accès gratuit à plus de 5 millions de documents.

Répondre aux questions suivantes à partir de l'annexe 4 et de la vidéo suivante (Gallica en vidéo) : *https://www.bnf.fr/fr/gallica-la-bibliotheque-numerique-de-la-bnf-et-de-ses-partenaires#bnf--gallica-en-vid-o-*

21. **Indiquer comment les documents numériques sont rendus accessibles ?**

La BnF a choisi la diffusion des ressources avec les formats suivants : JPEG2000, PDF, HTML, MP3, ePUB. À l'aide du tableau ci-dessous :

22. Rechercher les raisons de ce choix en mettant en avant les particularités de ces formats.

Format	Définition - particularités	Raisons de ce choix
JPEG2000		
PDF		
HTML		
MP3		
ePub		

23. Que signifie utiliser l'application en « marque blanche » ? Qu'est-ce que cela apporte aux partenaires de la BnF ?

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* utilise les outils du Web sémantique et s'inscrit dans une démarche d'ouverture des données. Mis en ligne en juillet 2011, data.bnf.fr continue d'évoluer et de s'accroître.

Après lecture de l'annexe 5 et en faisant des recherches sur Internet :

24. Définir ce qu'est une donnée publique ?

À partir de la page ci-dessous et d'éventuelles recherches complémentaires
http://www.univ-bpclermont.fr/root/Ressources_Num/Les_reseaux_sociaux_web_web/co/1-3_Web3.html

25. Définir et expliquer la notion de web sémantique.

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

26. Procéder de la manière suivante :

26.1. Saisir « Harry Potter » dans la barre de recherche.

26.2. Cliquer sur le premier lien obtenu dans la rubrique « Œuvres ».

26.3. Cliquer en bas de page sur télécharger en JSON.

26.4. Expliquer ce format de données et son utilité.

26.5. Citer 3 métadonnées obtenues. Quel est l'intérêt des métadonnées obtenues ?

À partir de l'utilisation de la carte mondiale dynamique du portail *https://data.bnf.fr/* :

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

28. Expliquer en quoi cet outil illustre bien la notion de big data ?

À l'aide de recherches sur Internet,

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

Après lecture de l'annexe 6 et en faisant des recherches sur Internet :

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

À partir des articles ci-dessous :
https://www.lebigdata.fr/data-lake-definition
https://www.lemagit.fr/etude/Groupe-SeLoger-quand-RGPD-rime-avec-agilite

31. Citer des entreprises ayant mis en place un lac de données, et dans quel but.

Annexes

Annexe 1 : Du document papier au document numérique

Extraits du site *www.bnf.fr*.

La BnF numérise à ce jour 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 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.

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 (*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 (*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.

Annexe 2 : La reconnaissance d'écriture

Extraits du site : *https://dataanalyticspost.com/la-reconnaissance-decriture-manuscrite-de-nouvelles-applications-pour-un-des-plus-vieux-problemes-dia/*

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.

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.

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

Extraits du site *www.bnf.fr*.

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[1] 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.

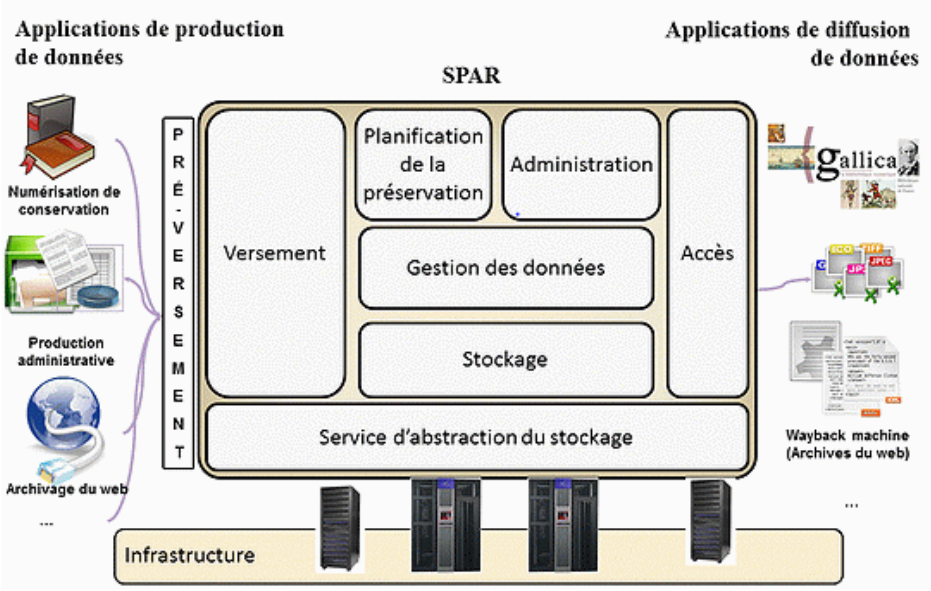


Schéma présentant le fonctionnement du magasin numérique Spar © BnF

[1] 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.

- 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.

  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).

Annexe 4 : L'accès aux documents numériques - la diffusion

Extraits du site *www.bnf.fr*.

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 ont été choisis par la BnF pour être le plus accessible 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 1663 vidéos. Un certain nombre d'ouvrages a fait l'objet d'une reconnaissance optique de caractères et le texte peut être recherché sur Gallica.

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. On peut citer entre autres :

- 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.

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

Extraits des sites : https://www.archimag.com/chiffre-du-jour/2015/11/26/big-data-aller-ou-pas
https://data.bnf.fr/

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.

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 ! …

A 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 ».

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.

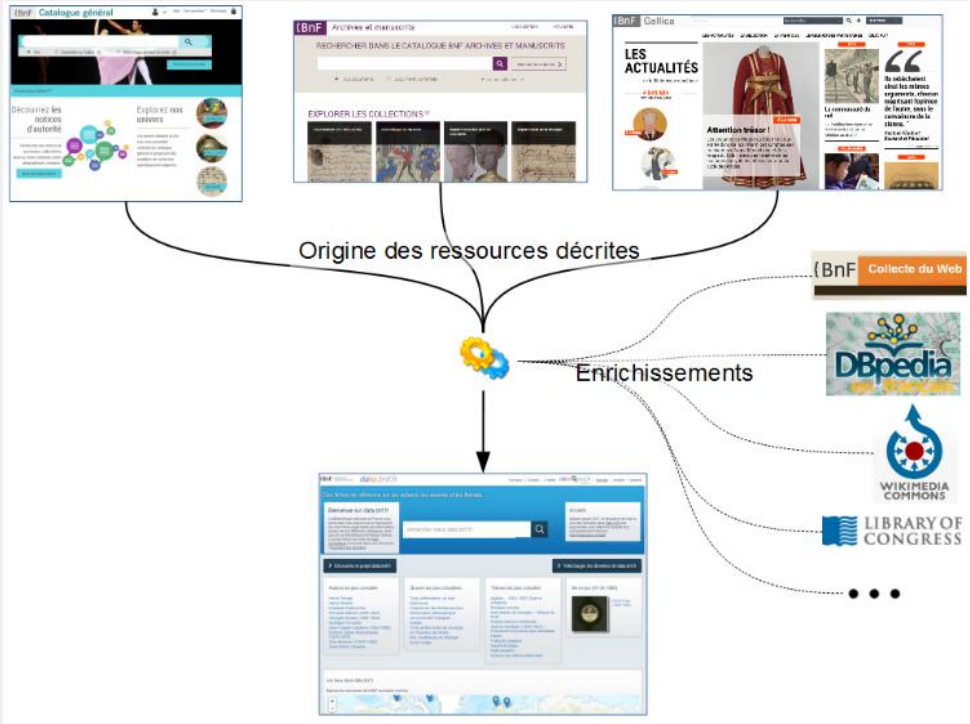
L'objectif est donc de valoriser la richesse des fonds de la BnF sur le Web et de servir de pivot entre les différentes ressources : *data.bnf.fr* est donc au service des autres applications de la BnF. Enfin, le projet s'inscrit dans une démarche d'ouverture de la BnF au Web de données et d'adoption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

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.

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'autorités 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.

Annexe 6 : *Big data* – lac de données

Extraits du site : *https://www.didaktic.fr/big-data/archivage_web/2/*

Actuellement, l'INA (Institut national de l'audiovisuel) s'engage dans une démarche de stockage des données orientée *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 inclue 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 bruts rendus 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.