Rapport "Bibliothèque numérique européenne" remis le 31 janvier 2006 au Ministre de la Culture.

Alexandre Moatti, Valérie Tesnière, Noémie Lesquins

BIBLIOTHEQUE NUMERIQUE EUROPEENNE

Rapport de synthèse et plan d'actions

Auteurs :

Alexandre Moatti, Ingénieur en chef des Mines, Secrétaire général du Comité de pilotage.

Valérie Tesnière, Conservateur général des bibliothèques, Secrétaire général adjoint du Comité de pilotage.

Noémie Lesquins, Conservateur des bibliothèques, Secrétariat général du Comité de pilotage.

BNUE Contexte 1 : Une course aux contenus de qualité et à leur indexation est lancée entre les 4 à 5 acteurs majeurs de l'Internet.

Depuis septembre 2005, de nombreux projets et accords concurrents de Google Print ont vu le jour :

- OCA Open Content Alliance fin septembre (dont Yahoo)
- Accord MSN-British Library début novembre.
- Volltextsuche-online éditeurs allemands novembre. NB : « volltextsuche »= recherche plein texte.
- Accord Google- American Library of Congress fin novembre « World Digital Library »

On constate au passage que le service Google Book Search est loin des résultats annoncés au point de vue des quantités (de l'ordre de 20 000 livres en un an), et que 70% des contenus sont des contenus provenant des éditeurs (contenus numériques natifs).

Les accords passés avec des bibliothèques sont un élément nouveau : les acteurs majeurs de l'Internet, après un temps de latence entre décembre 2004 et octobre 2005, engagent entre eux une lutte pour la numérisation et surtout l'indexation de contenus de qualité sur Internet . Le nombre de pages sur Internet est en effet le fonds de commerce de ces acteurs : s'apercevant que le nouveau contenu marginal sur Internet est maintenant de peu d'intérêt (blog,…), ils vont chercher et indexer eux-mêmes le contenu de qualité là où il est, dans les bibliothèques et chez les éditeurs.

La BNUE doit indiscutablement se positionner par rapport à ce nouveau contexte.

Fiche 1 Analyse des services lancés par les moteurs de recherche
Fiche 2 Une nouvelle stratégie des acteurs de l'Internet sur ces projets

BNUE Contexte 2 : L'effort français via Gallica : un acquis remarquable, mais une mutation importante est à présent nécessaire.

La France, avec Gallica et ses 80 000 documents, a été en avance. Mais, pour le public plus large visé par la BNUE, ou le public client des grands acteurs de l'Internet, ces documents ne correspondent pas aux standards de qualité visuelle sur Internet :

- Les documents sont en mode image, la recherche en mode texte est possible sur les tables des matières uniquement.
- Le format d'affichage est PDF (l'ouverture du logiciel Acrobat et la préparation du document prennent du temps).

- La qualité des images n'est plus dans les standards (marges noires, centrage de l'image sur l'écran, inclinaison de l'image,….)

Il est souhaitable de sortir de la logique d'une bibliothèque numérique faite pour des chercheurs : la BNUE doit aller vers des standards de visualisation conformes à ceux des moteurs de recherche Internet.

Fiche 3 Analyse de la bibliothèque numérique Gallica

BNUE Orientation proposée 1: une identité, une interface de recherche et un format de visualisation uniques

On consultera en annexe les différents sites étudiés (Google Print, Open Content Alliance contenus anglophones ; Persée, Cairn contenus de revues francophones).
Pour la BNUE, que ce soit pour les contenus patrimoniaux des bibliothèques ou les contenus payants en partenariat avec les éditeurs, il est proposé de retenir ce qui apparaît commun à ces différents sites :

- Une interface de recherche commune aux contenus patrimoniaux ou sous droits.
- Un format de visualisation unifié de contenus cherchables en plein texte.
- Un affichage des résultats de recherche en surbrillance.
- Un affichage page à page (ou deux pages par deux pages) en mode image PNG (Portable Network Graphics); le format PDF qui nécessite l'ouverture d'un logiciel à chaque nouveau document ne paraît pas le plus adapté pour la visualisation.

Il convient toutefois de laisser la porte BNUE ouverte à des contenus en mode image non OCRisables, aux conditions que ce soient des contenus pertinents, visualisables dans le même environnement que les autres (format de visualisation unifié), et qu'ils constituent une proportion limitée.

Fiche 4 Formats de visualisation BNUE

BNUE Orientation proposée 2: Choix d'architecture et schéma futur d'organisation.

Suite aux orientations du Comité de pilotage du 17 octobre, les groupes de travail reconstitués ont intensément travaillé sur les scénarios techniques possibles pour une BNUE :

- Architecture centralisée, de type base de données unique tenue par la BNUE.
- Architecture intermédiaire, de type Webservices avec OAI, avec identité et visualisation BNUE, les contenus restant chez leurs producteurs.
- Architecture décentralisée, de type OAI seul.

Le premier scénario correspond exactement au métier des moteurs de recherche, et à un investissement coûteux en termes de serveurs, de disponibilité,…Le troisième scénario est peu coûteux, mais ne préserve pas l'identité visuelle BNUE, qui devient un simple site-portail ouvrant vers d'autres sites.

Il sera difficile d'aller plus loin dans des comparaisons de coûts sans une étude de définition de besoins beaucoup plus précise, qui dépasse largement le cadre de ce rapport.

De manière pragmatique, l'effort de réalisation en phase 2 (mi-janvier à mi-juillet 2006) devra être porté sur le scénario intermédiaire (coût moindre au départ que scénario centralisé), et qui a l'avantage de respecter la visualisation et l'identité uniques BNUE.

Fiche 5 Architecture BNUE

Par ailleurs, nos groupes se sont penchés sur un schéma organisationnel possible de la BNUE à terme, définissant les fonctions de « *front-office* » (rapports avec les internautes et les hébergeurs) et de « *back-office* » (rapports avec les numériseurs) ; ce schéma organisationnel permet d'envisager ce que pourrait être une BNUE *à terme.* Ces scénarios organisationnels ont été conçus de manière cohérente avec les scénarios techniques.

Fiche 6 Schéma organisationnel BNUE

BNUE Présentation de premières maquettes (Thomson et Isako)

Deux maquettes ont été présentées au COPIL du 11 janvier.

La première maquette (adresse provisoire www.bnue.org), réalisée par Thomson, vise à illustrer les points suivants :

- Interface de visualisation propre BNUE.
- Interface de recherche commune BNUE dans un échantillon de documents.
- Effet démonstratif de la maquette vis-à-vis des partenaires de contenu et des partenaires européens.

A ce stade cette maquette est faite sur un serveur unique avec duplication de contenus, elle ne vise pas à tester les architectures préconisées.

La deuxième maquette (adresse provisoire www.bibnum.org) , réalisée par Isako, est très limitée en nombre de contenus, elle vise à illustrer l'OCRisation de documents et les qualités d'image en provenance de supports différents : OCRisation à partir de documents numérisés depuis livre papier, depuis microfilm ou microfiche, contenu électronique natif en provenance d'un éditeur.

Fiche 7 Présentation des maquettes COPIL 11 janvier

Il est proposé quatre actions, nombre volontairement limité même si elles peuvent comporter des sous-actions, à poursuivre sur l'année 2006.

BNUE Action 1 : OCRisation de masse

Il est important en 2006 de rendre les contenus Gallica (a minima ceux qui peuvent l'être, soit 70 à 80% du fonds) conformes aux standards de visualisation BNUE, c'est

à dire les OCRiser. Une évaluation donne pour 2006 sur cette partie du fonds un coût de 3,6M€ (budget à dégager BnF en liaison avec des partenaires privés, ou budget courant de numérisation 2006 BnF à utiliser ainsi).

Ceci a pour conséquence logique que dès 2006, les numérisations Gallica sont à faire en mode texte/image combinés conformément aux standards BNUE et aux standards Internet.

Au-delà, en 2007, se pose la question de la numérisation de nouveaux contenus. Au regard des résultats quantitatifs et qualitatifs de Google, la notion de numérisation de masse est à relativiser.
Au cours du premier semestre 2006, aura lieu un test de numérisation intensive (l'appel d'offres a été lancé par la BnF pour réponse avant le 30 décembre 2005). En fonction des résultats de ce test, la question se pose de l'opportunité d'une numérisation d'environ 150 000 ouvrages/an pendant deux ans (budget annuel environ 7,5 M€) ; un meilleur équilibre de dépenses doit être trouvé entre la masse des contenus à numériser et l'ouverture à un plus large public de ces ressources (dépenses de promotion et de partenariat).

Fiche 8 OCRisation des contenus Gallica

BNUE Action 2 : Une structure privé-public associant les éditeurs.

Le projet BNUE est une occasion d'initier, dans une démarche associant le public et le privé, et de manière coordonnée, visible et portée politiquement, la mise en ligne sur Internet de contenus sous droits, à discrétion de chaque maison d'édition et de manière rémunérée, en accompagnement de contenus patrimoniaux. Cette action peut aussi avoir une vertu pédagogique vis-à-vis des internautes.

Une plateforme BNUE mutualisée, avec format de visualisation unifié, donnant accès d'une part à des contenus patrimoniaux, d'autre part à des contenus sous droits suivant des modes de rémunération à définir, serait un plus incontestable pour les acteurs publics, les acteurs privés, et les internautes.

Les chantiers opérationnels (présente action et action 3 ci-dessous) devraient être menés par une entité juridique légère et à action rapide, prolongeant le Comité de pilotage dans sa composition actuelle ; dans cet esprit le schéma suivant est proposé :

- Fondation à Conseil de surveillance (Comité de pilotage BNUE actuel, conseil scientifique et directoire http ://www.recherche.gouv.fr/fondation/modelestatutcs.rtf, en charge du site-portail BNUE (action 3 poursuite de la maquette/ 400 000 € MCC) et de la coopération avec les éditeurs (action 2)
- Partenariat public-privé, association des éditeurs et d'industriels à cette Fondation (participation financière 10 000 à 20 000€, et plus pour les mécènes)
- Définition progressive d'un modèle économique, partie rémunérée du portail (contenus sous droits environ 20% des contenus dans un premier temps).

Fiche 9 Partenariat d'accès avec les éditeurs
Fiche 10 Quelle structure de partenariat privé-public ?

BNUE Action 3 : nouvelle phase du site BNUE.

La structure BNUE (Fondation) avance en marchant et rend compte au COPIL (son Conseil de surveillance) toutes les six semaines, avec le plan de route suivant :

- Nouvelle phase avec architecture (pour la maquette ce n'était pas nécessaire ni possible) ; mise en œuvre de la fiche 4.
- Nouvelle phase avec services : recherche multilingue et sémantique.
  Un travail doit porter sur le moteur de recherche du portail (sémantique, multiliguisme), brique qui pour l'instant n'a pas été étudiée en détail car non réalisable dans le cadre d'une maquette non rémunérée.
- Nouvelle phase avec contenus plus larges : définir et intégrer des contenus Gallica (et autres contenus francophones ou non francophones) suivant les axes de la future BNUE.
  Cette négociation avec les fournisseurs de contenu image-texte permettra d'enrichir rapidement les contenus BNUE, et de nous mettre en position d'ouvrir un site avec une masse critique de contenus.
- Ouverture grand public du portail (juin 2006)

Fiche 11 Quel portail BNUE à mi-2006 ?

BNUE Action 4 : poursuite du portage au niveau européen.

L'action de portage au niveau européen est à poursuivre comme suit :

- Suite à donner à la réponse à l'appel à idées UE du 20 janvier 2006.
- Contacts BnF avec autres bibliothèque nationales européennes. La BnF est à présent à même de contacter ses homologues sur la base des orientations BNUE arrêtées par le Comité de pilotage (Livre Blanc et Rapport de synthèse- Plan d'actions).
  Une lettre du Président de la BnF est partie en ce sens en décembre 2005, et la coopération doit à présent démarrer, sur ces bases, avec au moins une Bibliothèque nationale en Europe, y compris le cas échéant en associant ses contenus à la maquette.
- A terme mise à l'étude d'une structure européenne : GEIE dépendant du programme Minerva, ou Fondation européenne pouvant être OCA Europe.
  A cet égard, un accord européen a minima, qui n'a d'intérêt que s'il est rapide, pourrait être de faire front commun face aux sollicitations des moteurs de recherche, et négocier avec eux de manière commune et concertée.

Fiche 12 Portage de l'action au niveau européen

Toutefois, le projet BNUE ne doit pas attendre le résultat de ces différents chantiers européens pour avancer en France : il y a une réelle urgence à la mise en valeur des contenus francophones suivant les standards Internet.

Les actions 2 et 3 se proposent de réaliser cet objectif, dans une approche *« bottom up »,* en partant du terrain, tout à fait complémentaire et à mener en parallèle de l'action 4, qui correspond à une approche *« top down »,* par en haut.

C'EST UNE BNUE PAR CHANTIERS PARALLELES (ACTIONS 1 A 4) QUE NOUS PROPOSONS AU COMITE DE PILOTAGE, ET A LA STRUCTURE QUI PRENDRA SA SUITE, DE FAIRE PROGRESSER EN 2006 SOUS SON EGIDE.

@@@@@@@

Addendum : deux points importants

Hors plan d'action, nous souhaitons sensibiliser le Comité de pilotage sur deux sujets subsidiaires, mais structurants pour le projet BNUE, dans son volet francophone comme dans son volet européen.

- La nécessaire coordination des actions et financements sur la numérisation et les bibliothèques numériques, en France comme en Europe.

Fiche 13 Une coordination nécessaire des actions et des financements, en France et en Europe

- deux sujets juridico-techniques fondamentaux à long terme, mais dont il est nécessaire de se soucier dès à présent : la préservation des contenus numériques, et le dépôt légal sous format numérique des livres.

Fiche 14 Deux sujets complémentaires d'importance

Fiche 1 : Analyse des services lancés par les moteurs de recherche

L'objectif n'est pas de décrire ici l'intégralité des services lancés, mais de donner un éclairage sur les points intéressant la problématique BNUE (on trouvera dans le Livre Blanc une description et un historique détaillés).
On pourra se reporter par ailleurs aux annexes aux rapports BNUE pour examiner ces différents services analysés du point de vue de la visualisation.

Analyse des contenus Google Book Search

Le service Google Print a été rebaptisé Google Book Search (GBS) en novembre 2005. Il est composé de deux programmes distincts, pour les éditeurs et pour les bibliothèques.

> *L'articulation entre ces deux programmes est donné par la phrase suivante, extraite du site GBS :*
>
> *« Si vous êtes déjà un partenaire Google Livres, il vous suffit de nous indiquer les livres à intégrer dans votre compte en les ajoutant à votre liste de livres. Une fois ces livres référencés dans notre système, ils apparaîtront dans votre compte quand nous les aurons numérisés par le biais d'une bibliothèque.* ***Notez que notre Projet Bibliothèque englobe des fonds composés de millions de livres et que l'opération de numérisation devrait donc prendre des années. Nous numérisons les livres en suivant leur ordre de rangement****, et nous ne pouvons garantir la date à laquelle nous traiterons un livre en particulier. »*
>
> *(source :http ://books.google.com/intl/fr/googlebooks/publisher_library.html, « Informations destinées aux éditeurs concernant le Projet Bibliothèque »)*
>
> (les mises de texte en gras sont de nous)

Sur GBS, la recherche de livres est commune aux deux programmes. Nous avons fait des statistiques par année (recherche possible par année d'édition) : entre 1945 et 2005, on trouve une moyenne de 200 livres/an, principalement des livres sous droits du programme Editeurs. Ce sont environ 12 000 livres sous droits à fin décembre 2005 sur GBS [y compris sur les cinq dernières années, à titre indicatif on donne les chiffres : 227 (2005), 116 (2004), 158 (2003), 165 (2002), 192 (2001)].

Des exemples d'importants contributeurs éditeurs américains sont les suivants :
- O'Reilly Editions : 283 livres, informatique principalement.
- Arcadia Publishing : 300 livres.
- Blackwell Publishing : 290 livres.

- Stanford University Press 100 livres (ce ne sont pas les contenus patrimoniaux de la Bibliothèque de Stanford)

Du point de vue des éditeurs francophones, il a paru intéressant de regarder quels contenus étaient en ligne sur ce service à fin décembre 2005.
Les livres francophones sous droits proviennent des éditeurs[1] suivants :

- De Boeck Université (B), 300 livres.
- Editions de l'Eclat (F), 90 livres.
- Brill Academic Publishers (NL), *nombre de livres francophones non connu car plusieurs langues (270 livres au total)*
- Peeters Publishing (NL), *nombre de livres francophones non connu car plusieurs langues (165 livres au total).*
- Rodopi (NL), *nombre de livres francophones non connu car plusieurs langues (265 livres au total).*
- Editions « Alors hors du temps » (Marseille), 5 livres.

@@@@@@@@

Le service GBS est « localisé », comme beaucoup d'autres sites Internet : vous êtes automatiquement dirigé vers books.google.fr quand vous consultez depuis la France ; les possibilités d'achat de livres se font sur le site de l'éditeur (cf. ci-dessus) et les sites des trois grands vendeurs français en ligne : Amazon.fr, Fnac, Alapage (groupe France-Télécom).

Le service GBS est automatisé pour les éditeurs, qui transfèrent automatiquement leurs livres sous format PDF et les images de couverture sous format JPG avec un utilitaire (« Google Uploader »).
Sur la page d'aide aux éditeurs sur les chargements amont (https ://books.google.fr/partner/booklist-upload-windows) on peut voir l'écran :

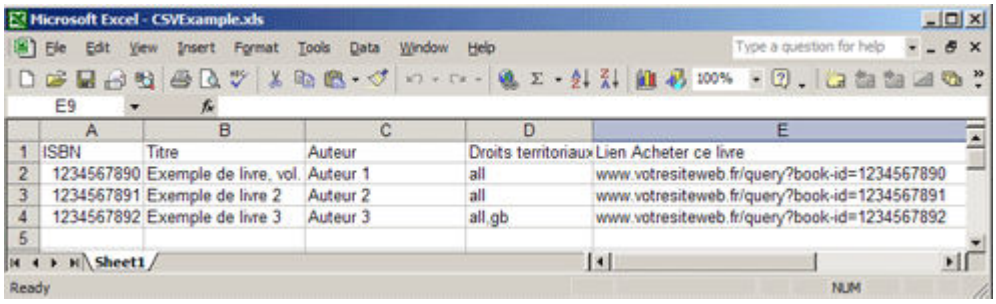


Figure 1.1 : Tableau à renseigner par l'éditeur ; on note le lien (facultatif) vers la page « achat du livre » sur son site ; on note aussi la rubrique « droits territoriaux », all.gb signifie que l'éditeur a les droits sur tous pays sauf GB (ce qui confirme que le service GBS est localisé).

Le contrat en ligne avec les éditeurs (voir https ://books.google.fr/partner/terms) est aussi instructif, sur deux points notamment :

[1] pour connaître les livres publiés sur GBS par un éditeur donné, http://books.google.fr/advanced_book_search?hl=fr&ie=UTF-8

- Google fait ses « meilleurs efforts » possibles pour limiter la consultation des ouvrages à 5-6 pages par ouvrage.
- L'éditeur doit accepter un système de publicités sur les pages de ses livres, système pour lequel il est rémunéré.

Apparition de publicités non liées aux livres en bas des pages Google Book Search :

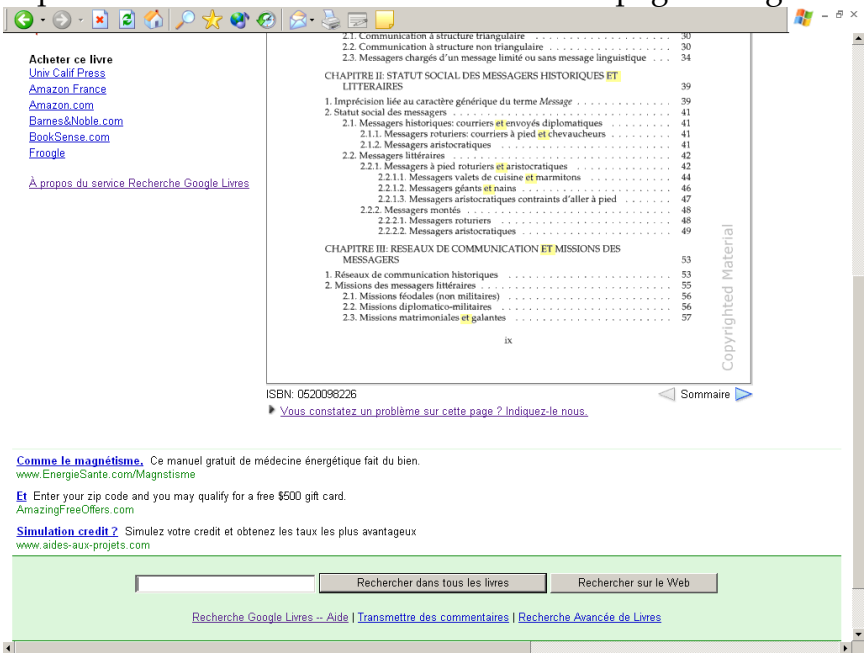


Figure 1.2 : en bas d'un livre en français (revue fournie par University of California Press), on voit apparaître des publicités : « Comme le magnétisme », « Simulation crédit ? » Il n'y a de publicité que sur les pages GBS « Editeurs », pas sur les pages GBS « Bibliothèques ».

@@@@@@@

Il est beaucoup plus difficile de faire une étude quantitative sur les contenus patrimoniaux (Programmes Bibliothèques), puisqu'il n'y a pas d'outil de recherche spécifique sur ce programme, qui semble-t-il s'est développé moins vite que les annonces quantitatives de décembre 2004 ne le laissaient penser :

« Pour le moment, la plupart des livres viennent du Programme Partenaires Google Livres, un programme Web qui permet aux éditeurs de toutes tailles d'inclure leurs ouvrages dans les résultats de recherche de Google. (...) Nous avons également mis au point des partenariats avec certaines grandes bibliothèques dont les livres devraient progressivement apparaître dans les résultats des recherches effectuées sur Google. »
(extrait de http ://books.google.com/intl/fr/googlebooks/help.html#8)

A janvier 2006, on peut dénombrer ainsi ce programme :
- Avant 1700 non significatif (quelques unités par années)
- XVIII°s : 1 013 ouvrages.
- 1800-1860 : environ 75 livres/an, soit 4 500 livres (ils sont numérisés par Google).
- 1860-1930 : en cours de numérisation semble-t-il, on ne voit que la couverture du livre. En revanche, avant 1860, les ouvrages apparaissent intégralement (avec la mention « numérisés par Google »)

Un nombre non négligeable d'ouvrages francophones (30% environ) sont parmi ces livres, numérisation d'ouvrages français appartenant aux bibliothèques américaines.

Par rapport aux annonces faites en décembre 2004, du point de vue quantitatif treize mois plus tard, le service GBS se compose d'au maximum 20 000 livres dont 70% provient des éditeurs (une centaine d'éditeurs), sur la base d'un contenu électronique natif. Le Programme Bibliothèque, qui va de pair avec la numérisation, semble démarrer plus lentement.

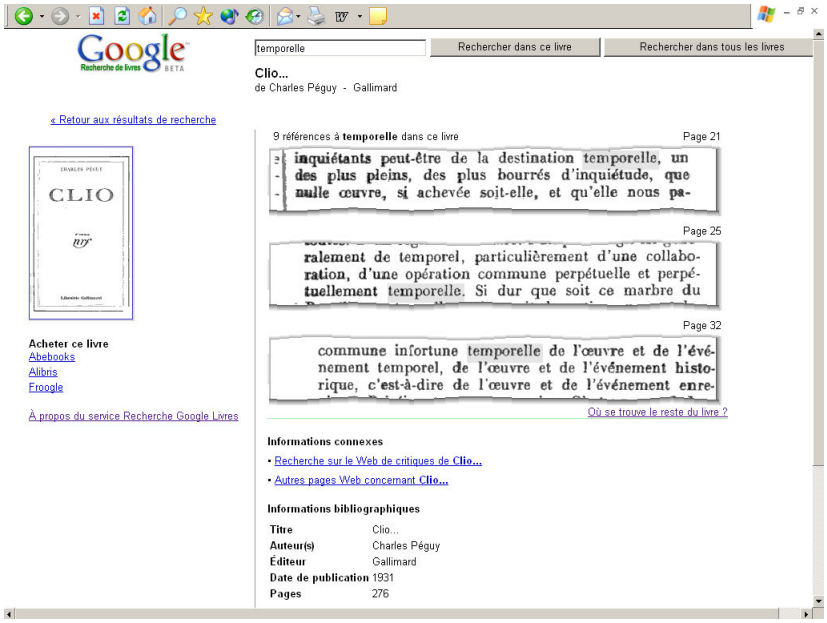


Figure 1.3 : Livre issu du Programme Bibliothèque. Livre français, édition Gallimard 1931 de Péguy. Livre sans doute dans une bibliothèque américaine, en cours de numérisation. Aucun contenu intéressant sur ces livres qui apparaissent pour l'instant ainsi découpés.

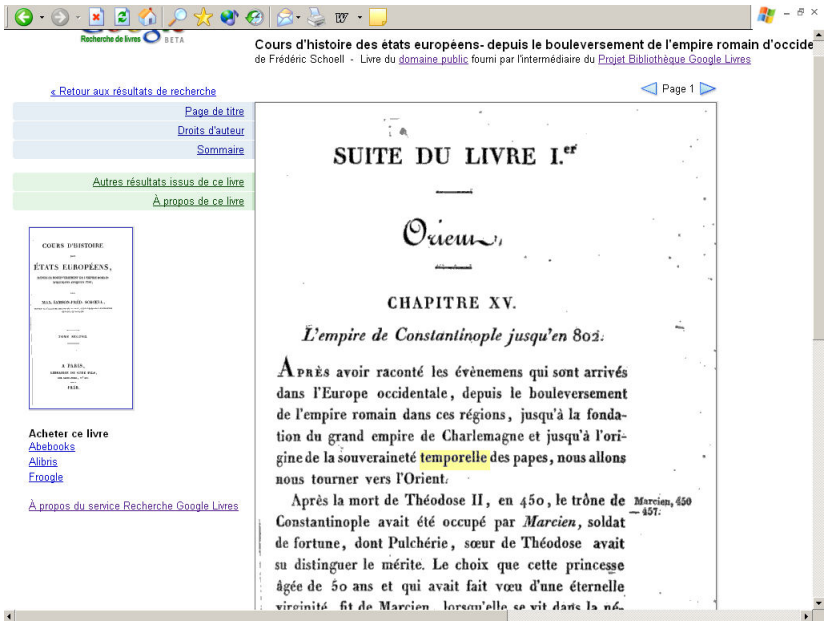


Figure 1.4 : Livre issu du Programme Bibliothèque comme indiqué. Livre sans doute dans une bibliothèque américaine (date 1830). Consultable intégralement.

@@@@@@@

Enfin, il convient de signaler une qualité de numérisation parfois plus que déficiente sur certains ouvrages francophones, y compris contemporains.

Temps Et Aspect- De LA Grammaire Au Lexique
de Veronique Lagae - Language Arts & Disciplines - 2002
Page 121 - **...** a tine division **temporelle**: Las compliments de localisation **temporelle** sans pr/position 121 **...** denote tine notion **temporelle** ou encore qua son trait **...**
[Autres résultats issus de ce livre]

Figure 1.5 : Voir texte apparaissant dans les résultats de recherche ; « les compléments » deviennent « las compliments ».

Accord entre MSN et British Library

(source http ://www.bl.uk/news/2005/pressrelease20051104.html + réunion avec MSN le 17 novembre 2005)

Beaucoup plus que pour Google, les chiffres de l'accord sont clairs : il consiste à numériser et à indexer sur un an 25 millions de pages, soit 100 000 ouvrages libres de droits. Le coût de la numérisation, offert par Microsoft, est évalué à 2,5 Millions $, soit 10c. la page.

MSN, au cours de notre rencontre avec eux, a insisté sur le travail d'indexation de ces contenus qu'ils allaient conduire : c'est en effet un point crucial, y compris pour la BNUE.

Ces contenus seront visibles, entre autres, depuis le service MSN Book Search qu'ils comptent ouvrir en 2006.

Toutefois, il n'y a pas d'exclusivité, et n'importe quel moteur de recherche pourra indexer ces mêmes contenus une fois qu'ils seront sur Internet. Gageons simplement que MSN prend une longueur d'avance sur ses concurrents, la stratégie étant :

- Ouverture de MSN Book Search en 2006, le gros du contenu étant celui de la British Library (masse critique à l'ouverture).
- Avantage concurrentiel de « timing » sur les autres moteurs, ceux-ci ne pouvant indexer le contenu qu'une fois le service ouvert ; on peut aussi penser que MSN, concevant et réalisant la numérisation, aura plus de facilités techniques pour assurer l'indexation.

Le lien de recherche sera développé en liaison avec les contenus de l'encyclopédie MSN Encarta (une idée analogue a été développée en France avec le « Copilote » de Larousse). C'est une manière d'enrichir la recherche, et de faire connaître le produit Encarta pour le vendre.

Enfin, cet accord comporte, volet moins connu, une coopération de Microsoft (et pas uniquement MSN) avec des laboratoires de recherche de Cambridge University pour la préservation à long terme des contenus numériques.

Accord entre Library of Congress et Google

(source http ://www.loc.gov/today/pr/2005/05-250.html annonce du 22 novembre 2005)

Cette annonce nous interpelle, d'autant plus que la Library of Congress est un partenaire privilégié de la BnF dans la mise en commun de documents numériques sur Internet (voir présentation de la BnF au comité de pilotage du 30 août, et copies d'écran illustratives sur fiche 4 Architectures).

Les chiffres sont, à la différence de MSN – British Library, moins clairs. Il s'agit d'un programme de « mécénat », dont Google est le premier contributeur à hauteur de 3M$.

- La base de documents accessibles en ligne de la LoC s'élève actuellement à 10 millions de documents uniques (dont photos, cartes,…) dans le fonds « American Memory » visant à constituer une mémoire numérique de l 'histoire américaine.

Le concept de la « World Digital Library » (WDL), tel que proposé par le Directeur de la LoC à l'UNESCO (voir discours du 6 juin 2005 http ://www.loc.gov/about/welcome/speeches/wdl/wdl_6-6-05.html) est intéressant à plus d'un titre :

- C'est le prolongement direct du rapport du PITAC (Président's Information Technology Advisory Committee) de 2001 intitulé « *Digital Libraries : Universal Access to Human Knowledge* », et de la « *National Digital Library* » lancée par LoC dès 1995.
- Les coopérations entre la LoC et l'Espagne, les Pays-Bas, la France (BnF) y sont présentées comme visant à éclairer l'histoire coloniale des Etats-Unis…
- La perche est tendue vers l'UNESCO pour contribuer à la mise en place d'une WDL, prioritairement tournée vers « les cultures chinoises et d'extrême-orient, du subcontinent indien, du monde islamique de l'Indonésie à l'Afrique ». L'Europe semble ne pas y avoir sa place, autrement que pour éclairer l'histoire coloniale des Etats-Unis…

C'est aussi parce que la moitié des documents de la LoC sont dans une autre langue que l'anglais que la LoC est en mesure de proposer ce concept de WDL. On peut penser que les fonds de mécénat (dont Google) iront en priorité à la numérisation de ces documents, permettant de constituer des corpus numériques « utilisables par d'autres bibliothèques à travers le monde ».

Enfin, point à signaler, LoC a déjà effectué avec Google un test de numérisation de 5000 documents, test incluant la manipulation de documents fragiles.

Open Content Alliance (OCA)

(voir dépêche GFII 663 du 11 octobre, dossier Comité de pilotage du 17 octobre)
(www.opencontentalliance.org)

OCA se distingue de Google GBS par sa nature consortiale, fédérant plusieurs partenaires, industriels notamment :

- Le consortium OCA regroupe, à la différence de GBS, plusieurs industriels de premier rang : Yahoo !, Adobe, Hewlett-Packard, avec semble-t-il des rôles bien déterminés à chacun :
  - ✓ HP contribue au prêt de numériseurs de base dans les bibliothèques qui en sont démunies.

- ✓ Adobe pour les formats de numérisation (le format d'affichage n'est heureusement pas du PDF, mais reste à étudier, voir ci-dessous).
- ✓ Internet Archive organise des centres de numérisation dans certaines bibliothèques plus importantes et assure l'hébergement.
- ✓ Yahoo indexe les contenus.

➢ La recherche de livres ne se fera pas dans un environnement lié à une entreprise comme l'est GBS, mais dans un environnement ouvert.
On peut avoir une idée du site de consultation de livres d'OCA, hébergé par Internet Archive, à www.openlibrary.org ; il faut constater toutefois que ce site est pour l'instant une maquette assez anecdotique. Par ailleurs le format d'affichage ne paraît pas très fluide, et devrait être étudié techniquement plus avant (une image PDF créée à la volée dans un fichier PHP)

C'est bien la logique coopérative de partenariat entre le secteur public (Bibliothèques) et le secteur privé (Editeurs et gros industriels) qui est l'aspect innovant de cette alliance ; comme le conclut le GFII :
« Certaines bibliothèques voient dans l'OCA une plate-forme de collaboration permettant de construire des collections dans une logique de partage des ressources. Dans cette optique de partage, les budgets de numérisation pourraient ne plus être vus comme un coût additionnel mais comme une condition sine qua non pour pouvoir apporter des fonds servant de monnaie d'échange dans le cadre de ce partage programmé de ressources numérisées.
Là aussi il sera intéressant de vérifier à plus long terme si cette logique de coopération moins gourmande en investissements pourra générer des bibliothèques numériques capables de rivaliser avec celle que projette Google en misant sur son impressionnante force de frappe financière. »

Fiche 2 : Depuis octobre 2005, une nouvelle stratégie des grands acteurs de l'Internet sur ce type de projets.

Au cours des mois de travail du Comité de pilotage entre septembre et décembre, nous avons eu clairement l'impression d'une accélération de la stratégie des acteurs comme Yahoo et MSN à partir de la mi-octobre. Après une période d'attentisme et de latence après l'annonce de Google en décembre 2004, et les avatars connus par Google à l'été 2005 (procès de la Guilde des Auteurs), ces acteurs majeurs de l'Internet ont défini leur stratégie et la mettent en œuvre dans chaque pays.

De la même manière que Google a ouvert son site Google Book Search, MSN et Yahoo voudront ouvrir courant 2006 un site grand public analogue avec une masse critique de contenus :

- MSN ouvrira « *MSN Search Book* » (ou équivalent) en 2006 avec la masse critique de contenus British Library. Bien évidemment, une fois le service ouvert, les moteurs concurrents comme Yahoo ou Google ou AOL peuvent aussi indexer ces contenus, mais il y a un incontestable avantage de « timing » commercial, permettant d'ouvrir un service et d'en faire la publicité avec une masse critique de contenus.
- Yahoo cherchera aussi la masse critique d'une bibliothèque-phare pour ouvrir son service « *Yahoo Search Book* » en 2006 (le site Open Content Alliance ne peut être assimilé à un service grand public avec une masse critique).

Une accélération des contacts de la part des grands moteurs de recherche.

Nos premiers contacts avec Yahoo ! en septembre 2005 étaient assez lents, avec des réponses tardives. A partir de mi-octobre, et l'annonce OCA, nous avons été relancés comme nous l'avons été et le sommes par MSN suite à notre premier contact.
A ce jour, MSN et Yahoo sont les deux sociétés qui veulent conclure un accord avec la BNUE et/ou la BnF, du type de celui qui a été conclu avec la British Library. Ces sociétés ont notamment bien intégré dans leur stratégie l'avance de la France en terme de numérisation de contenus, et le rôle qu'elles pourraient jouer dans l'OCRisation de ces contenus déjà numérisés en mode image.

Nous pouvons penser que ces deux grands acteurs de l'Internet sont d'autant plus attirés par les contenus francophones, que ces contenus peuvent intéresser la population francophone nord-américaine (Canada), qui fait partie de leur marché domestique, où ils sont en proportion plus présents qu'en Europe.

A contrario, cela fait contraste avec les « moteurs de recherche européens » (type Exalead), qui n'ayant pas de stratégie grand public, ne sont pas engagés dans cette course aux contenus de qualité de l'Internet. Il est important, à cet égard, de bien

faire la différence entre des acteurs grand public de l'Internet qui se trouvent être des moteurs de recherche (Yahoo, Google, MSN), et des moteurs de recherche *stricto sensu* qui, souvent plus performants techniquement, ne visent pas une notoriété grand public car ayant d'autres objectifs (Exalead, Quaero,…) ; ces derniers pourraient cependant être des partenaires techniques de grande qualité pour le moteur BNUE lui-même.

Une stratégie alternative d'association de partenaires européens au projet BNUE a été évoquée dans un groupe de travail COPIL : elle vise à faire des fournisseurs d'accès Internet, qui eux sont tous des groupes européens, des partenaires d'accès du site BNUE (fiche 10).
Elle ne saurait remplacer une stratégie de partenariat avec un moteur grand public d'indexation comme MSN ou Yahoo, mais peut être vue comme complémentaire, en répartissant le partenariat d'accès grand public aussi sur des groupes européens.

Fiche 3 : analyse de la collection numérique Gallica

Avec la collection numérique Gallica http ://gallica.bnf.fr, notre pays a incontestablement pris une longueur d'avance en ce qui concerne les contenus numériques accessibles via Internet.

Rappel sur la collection Gallica

Actuellement ce sont environ 80 000 volumes sur Internet, dont 80% de monographies et 20% de revues.
La répartition par périodes est la suivante : Antiquité 3%, Moyen-Age 5%, XVIe siècle 7%, XVIIe siècle 6%, XVIIIe siècle 21%, XIXe siècle 40%, Xxe siècle 18%. La proportion plus faible d'œuvres du XX° siècles est bien évidemment due à ce que la plupart de ces œuvres (jusqu'à 1935) sont encore sous droits.

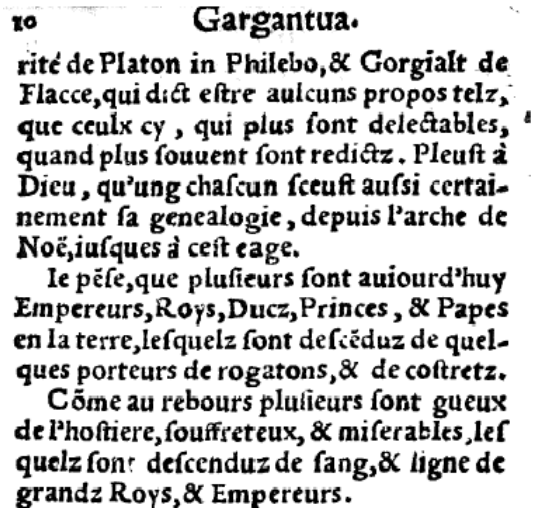

10 Gargantua.
rité de Platon in Philebo,& Gorgialt de Flacce,qui dict eſtre aulcuns propos telz, que ceulx cy , qui plus ſont delectables, quand plus ſouuent ſont redictz. Pleuſt à Dieu , qu'ung chaſcun ſceuſt auſsi certainement ſa genealogie , depuis l'arche de Noë,iuſques à ceſt eage.
Ie pẽſe,que pluſieurs ſont auiourd'huy Empereurs,Roys,Ducz,Princes , & Papes en la terre,leſquelz ſont deſcẽduz de quelques porteurs de rogatons,& de coſtretz.
Cõme au rebours pluſieurs ſont gueux de l'hoſtiere,ſouffreteux, & miſerables,leſ quelz ſont deſcenduz de ſang,& ligne de grandz Roys,& Empereurs.

Figure 3.1 : exemple d'une édition Gallica de Gargantua (Rabelais 1542) ; à titre d'exemple, la collection Gallica ne contient pas de texte de Gargantua en français moderne (il existe deux autres éditions numérisées du XIX°s, mais toujours en français ancien).

Gallica connaît actuellement un succès en croissance : 180 000 connexions par mois, et environ 1 500 000 consultations de documents par mois ; 70% des visiteurs sont français, et la moitié des visiteurs étrangers sont francophones (Gallica est un élément de rayonnement de la francophonie).

Les modalités de la numérisation

Il est important d'exposer ici les principaux modes de numérisation, pris par exemple sur l'année courante 2005 (1, 8 millions de pages soit 6 000 ouvrages environ) :

1. Numérisation directe avec destruction du livre : 55%
2. Numérisation directe sans destruction du livre : 10%
3. Numérisation à partir de microfiches ou microfilms : 35%.

Concernant le mode 3, il se base sur le fait qu'un stock important de microformes (fiches ou films) de bonne qualité avait été constitué à la BnF dans les années 1960 pour assurer la conservation de certains papiers de mauvaise qualité du XIX° siècle.
Concernant le mode 2, il est actuellement en croissance.
Concernant le mode 1, il s'applique principalement, là aussi, au XIX° siècle, car en-deçà, le coût d'achat de doubles est important. Il convient toutefois de signaler que le budget dit « d'antiquariat » (correspondant à l'achat de doubles à détruire dans une numérisation) est limité et en diminution, de l'ordre de 30 à 40 000€/an.

Une petite partie de documents en mode texte

Environ 1,5% des ouvrages numérisés sur Gallica (environ 1250 ouvrages) le sont en mode texte, le reste l'est en mode image. Les fichiers Gallica en mode texte sont de deux sources :

- 1100 textes « Frantext » numérisés par le CNRS (Institut National de la Langue Française InaLF) ; ces textes ne sont parfois que des fragments et sont difficilement exploitables.

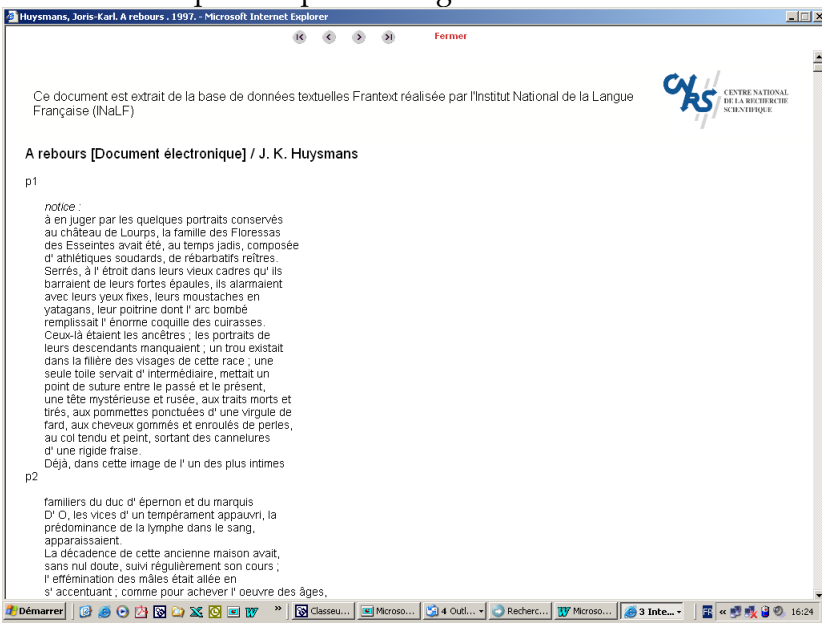


Figure 3.2 : œuvres littéraires en mode texte Gallica/ Frantext

- 140 textes de Chateaubriand et Balzac, complets, numérisés en partenariat avec les éditions Acamedia (aujourd'hui disparues) ; textes de même provenance qu'on retrouve d'ailleurs dans de nombreux sites de diffusion culturelle.

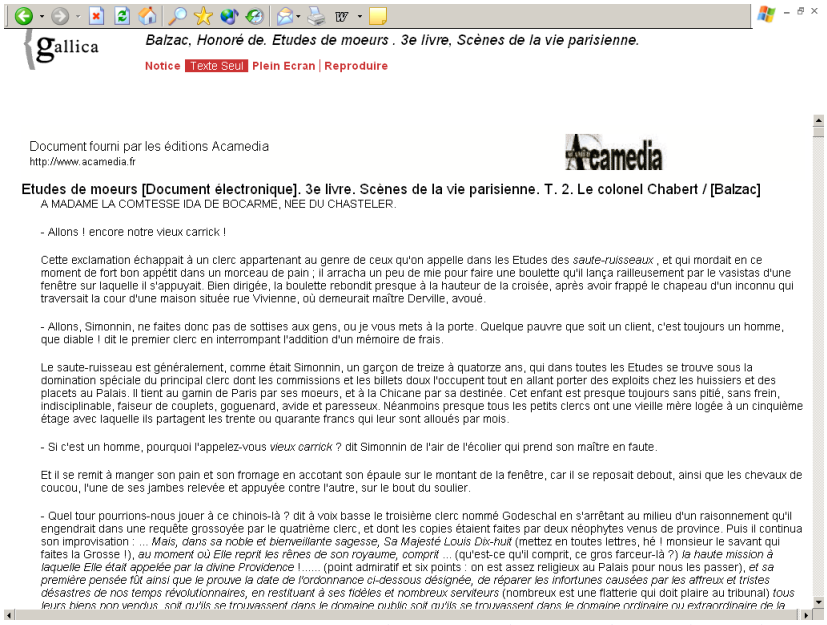


Figure 3.3 : œuvres en mode texte Balzac et Chateaubriand

Dans ces documents, la recherche plein-texte est possible. Il n'y a pas de visualisation en mode image pour ces documents (c'est une longue page HTML qui défile).

Numérisation en mode image et visualisation

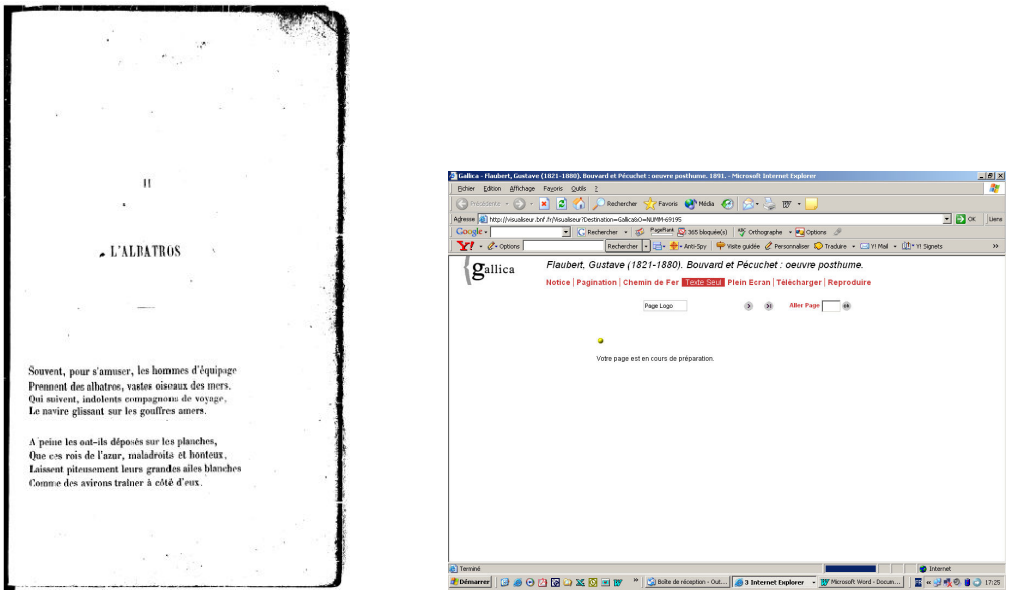
L'immense majorité des documents Gallica est, pour des raisons de budget de numérisation, visible en mode image.

La recherche plein-texte est possible, pour une partie des documents Gallica, sur la table des matières, ce qui est intéressant pour certains ouvrages et pour certains publics.

Mais sur le contenu des documents, aucune recherche plein-texte n'est possible ; par ailleurs le format de visualisation est le PDF, ce qui fait qu'il y a deux étapes préalables :

- ouverture du logiciel Acrobat sur l'ordinateur de l'internaute, au début de consultation d'un document.
- préparation de la page demandée (encapsulage du fichier TIFF de conservation vers le fichier PDF de visualisation : message « *votre page est en cours de préparation* »)

Par ailleurs les numérisations en mode image ont été faites sans correction d'images (redressement d'une image, centrage, effacement des marges noires,…), souvent à partir de microformes (35%), ce qui conduit à une esthétique de visualisation assez éloignée des produits lancés actuellement par les grands moteurs.



II

L'ALBATROS

Souvent, pour s'amuser, les hommes d'équipage
Prennent des albatros, vastes oiseaux des mers,
Qui suivent, indolents compagnons de voyage,
Le navire glissant sur les gouffres amers.

A peine les ont-ils déposés sur les planches,
Que ces rois de l'azur, maladroits et honteux,
Laissent piteusement leurs grandes ailes blanches
Comme des avirons traîner à côté d'eux.

Figure 3.4 : à gauche, poème des Fleurs du Mal de Ch. Baudelaire (marques du livre) ; à droite, message d'attente « votre page est en cours de préparation ».

Deux autres points à signaler.

- Un premier point est l'indexation par les moteurs de recherche. Ceux-ci n'indexent pas les contenus produits à la volée, sans adresse Internet permanente (contenus de type PHP, JSP, Flash, contenus PDF créés à la volée à partir de TIFF comme Gallica…). Si les contenus Gallica sont mieux indexés depuis quelque temps, c'est grâce au protocole OAI (voir fiche 4) qui indexe les métadonnées des notices.
- Un deuxième point est que les fichiers en mode image seul (PDF mode image) sont totalement inaccessibles aux handicapés déficients visuels.

Facteurs à considérer d'évolution de la collection Gallica.

Deux facteurs sont à considérer dans l'évolution technique de Gallica.

Le premier est déjà engagé et consiste à rendre compatibles au protocole OAI les descriptifs des ouvrages (« métadonnées descriptives »), ce qui dans un premier temps pallie l'absence de mode texte, et permet à un moteur de recherche grand public, ou à un moteur spécialisé (« moissonneur OAI ») d'indexer le contenu des ces notices, à défaut d'être en mesure d'indexer le texte complet de l'ouvrage.
D'ores et déjà un effort significatif a été engagé depuis 2004 par la BnF sur le sujet, et ce sont actuellement 24 000 notices Gallica (soit toutes les monographies en un volume), qui bénéficient d'une notice OAI. Ceci permet bien entendu une meilleure visibilité de ces contenus sur le Web.

Le deuxième facteur d'évolution à considérer est la mise à niveau en mode texte de l'intégralité des contenus Gallica (« OCRisation »).

Un premier test a été effectué dans le cadre de la maquette BNUE, avec l'OCRisation d'un ouvrage numérisé en mode image dans Gallica :

Figure 3.5 : OCRisation d'un document Gallica (Henri Poincaré, « La science et l'hypothèse »)

Une projection d'action « OCRisation de masse » des contenus Gallica peut être réalisée sommairement comme suit :

- masse visée : 80% des contenus Gallica, XVIII° siècle et au-delà, soit environ 65 000 ouvrages.
- Coût à la page 0,20€ : ce coût comprend une amélioration de la qualité de l'image telle qu'elle figure sur Gallica : redressement de la page, centrage de la page par rapport à l'écran, suppression des marques noires sur les bords de l'image (correspondant à la trace du livre lors de la numérisation).

Une conséquence importante est à tirer de cette évolution, si elle est jugée nécessaire : les nouvelles numérisations dans la collection Gallica à partir de 2006 devraient se faire en mode texte.

Fiche 4 : Format de visualisation BNUE

Format de visualisation proposé pour la BNUE.

On consultera en annexe 1 les différents formats de visualisation étudiés (Google Print, Open Content Alliance contenus anglophones ; Persée, Cairn contenus de revues francophones).

Pour la BNUE il est proposé de retenir ce qui apparaît commun à ces différents formats :

- Une visualisation unifiée de contenus cherchables en plein texte.
- Un affichage des résultats de recherche en surbrillance.
- Un affichage page à page (ou deux pages par deux pages) en mode image PNG (Portable Network Graphics)[2] ; le format PDF qui nécessite l'ouverture d'un logiciel à chaque page est à proscrire pour la visualisation.

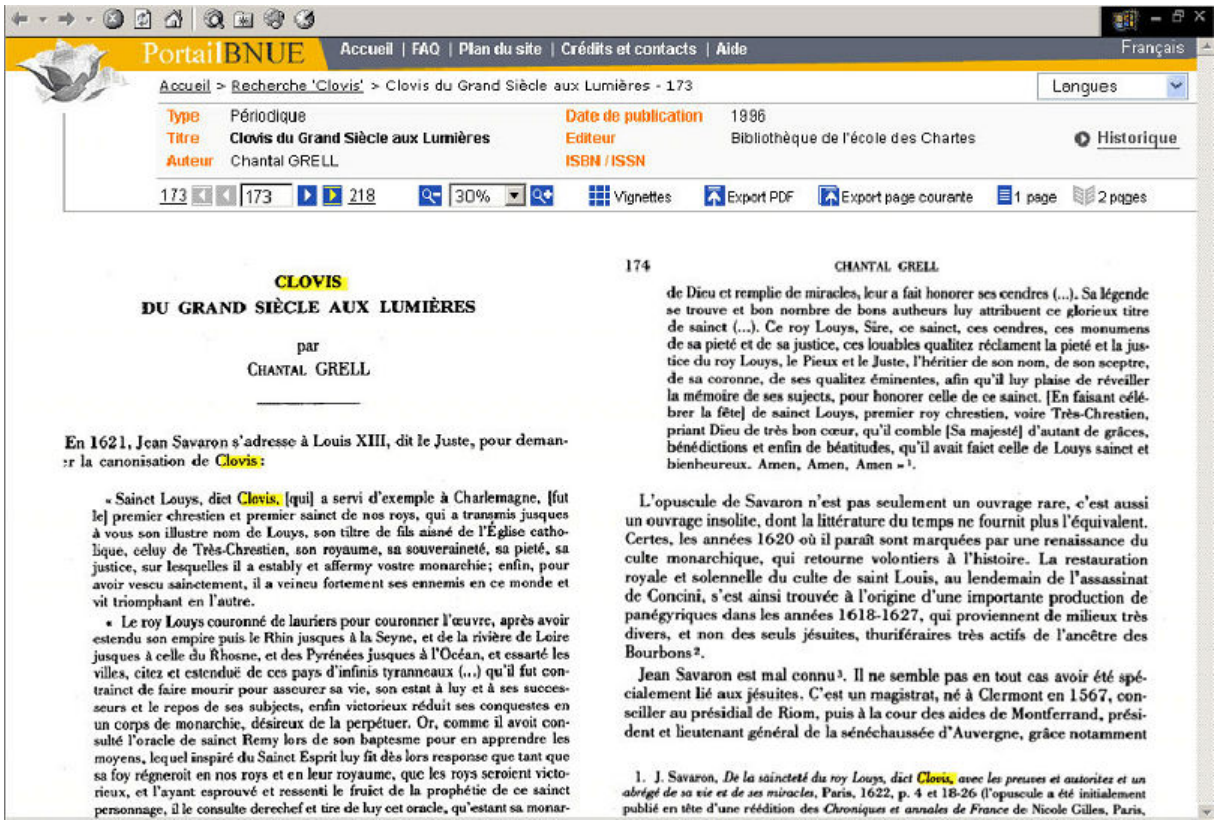


Figure 4.1 : Maquette BNUE, exemple de visualisation sous format unique, modes image et texte combinés, affichage en surbrillance des résultats de recherche.

L'association des modes image et texte semble être le choix de visualisation à présent couramment répandu, qu'il soit fait par des acteurs majeurs de l'Internet (Google dans Google Print, Yahoo dans Open Content Alliance), ou qu'il soit fait par des

[2] Le format PNG est un format graphique de type « bitmap », créé pour offrir une alternative libre au format graphique GIF, qui est un format propriétaire. Il est largement répandu sur Internet.

acteurs documentaires beaucoup plus spécifiques (sites français Persée, Cairn, Numdam,…).

Avantages des modes image et texte combinés.

Les avantages du mode image sont les suivants :

1. Il permet de donner à l'internaute l'impression d'un livre, de le feuilleter page à page.
2. Il permet aussi de garantir l'authenticité de la source : grâce aux premières et dernières images (couverture, page de garde, 4° de couverture, notices légales d'imprimeur avec date d'impression, le cas échéant table des matières,…), l'internaute peut vérifier qu'il est bien dans un livre donné.

Ceci peut paraître tautologique, mais le mode « texte + image » combine les avantages des deux modes : les deux avantages 1. et 2. ci-dessus sont importants par rapport à un simple fichier en mode texte, où la plupart du temps il n'y a pas « feuilletage » (on déroule une longue page HTML), et où on ne peut s'assurer du contenu d'aucune manière (voir fiche 3, figure 3.2 et 3.3)

Les avantages du mode texte subsistent bien évidemment :

3. Il permet une souplesse d'accès au document : l'internaute n'est pas obligé de voir son logiciel Acrobat Reader s'ouvrir à chaque consultation de nouveau document, le document est par ailleurs moins lourd → le temps d'affichage en est réduit d'autant.
4. Il permet une recherche plein texte, et par voie de conséquence une indexation par les moteurs de recherche → les contenus sont beaucoup plus visibles sur Internet.
5. Il permet éventuellement dans une phase ultérieure le retraitement du texte en vue de la traduction (ce qui est important pour le projet BNUE), la transcription et l'accessibilité pour les publics empêchés.

Il a paru aux membres des groupes de travail BNUE que le format de données XML dit « ALTO »[3] (*Analyzed Layout and Text Object*) est très intéressant en tant que format spécialement dédié au traitement d'une version textuelle obtenue automatiquement (donc peu coûteuse) en rapport avec l'image à laquelle elle correspond. C'est un standard ouvert qui pourrait être aisément adopté par toutes les institutions désireuses de faire du traitement automatisé de textes. Le fait de promouvoir de tels standards qui permettent de gérer à la fois le texte et l'image pour un document numérisé et qui ouvrent des possibilités intéressantes en matière de traitement automatique des données favorisera à coup sûr des initiatives de numérisation plus vastes et plus accessibles.

[3] Le format ALTO est développé par la Library of Congress comme extension d'autres formats, voir feuille de style-type à http://www.loc.gov/ndnp/alto_1-1-041.xsd

Conséquences sur les contenus de la BNUE.

Une conséquence importante est à déduire de ce choix quant aux contenus de la BNUE ; elle exclut, au moins dans un premier temps, les documents non cherchables en plein-texte :

- Documents numérisés en-deçà de 1800 : compte tenu de la qualité de l'impression, de la présence de caractères anciens, de phonèmes anciens, l'OCR n'est pas valable sur ces documents (à titre d'exemple, 42% des ressources Gallica sont en-deçà de 1800).
- Documents manuscrits (de type archives), et cartes.

Il est toutefois proposé d'accepter dans la BNUE des contenus non OCRisables, s'ils sont d'un grand intérêt pour la BNUE, et à la double condition suivante :

- qu'ils soient visualisables sur le site BNUE de la même manière que les autres documents (visualisation unifiée) (p.e. imprimés antérieurs à 1800, livres manuscrits, archives).
- qu'ils ne représentent pas plus d'un certain pourcentage, par exemple 20%, des contenus de la BNUE.

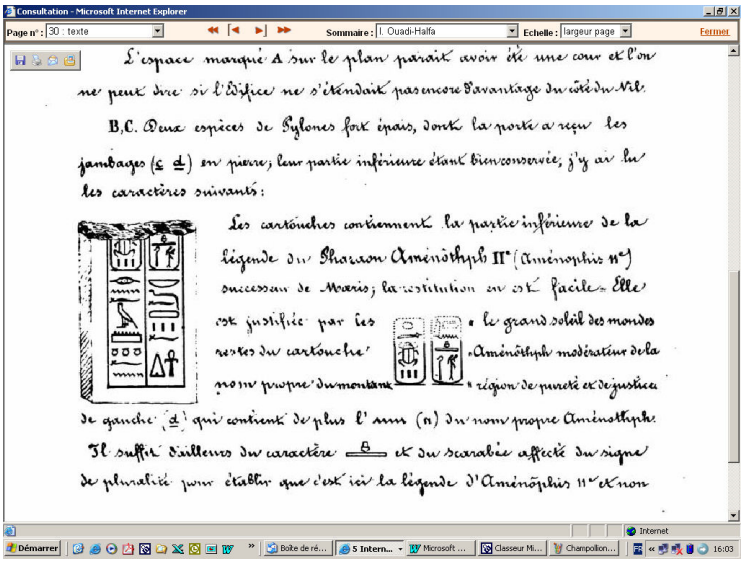


Figure 4.2, livre manuscrit de Champollion, site Maison de l'Orient et de la Méditerranée (MOM). Ce type de contenus pourrait être intégrés sous visualisation BNUE sous la double condition ci-dessus.

Fiche 5 : Architectures techniques possibles BNUE

Les choix structurants pour la BNUE ne peuvent se faire sans examen des différents scénarios d'architecture possibles.

Les sous-groupes de travail BNUE ont étudié ces scénarios de manière intensive, et grâce à l'appui actif de la Direction des Archives de France, nous avons pu trouver un langage commun entre divers interlocuteurs sur les différents choix possibles, langage commun qui sera important pour la suite du projet BNUE.

Du point de vue des coûts, il est difficile d'aller plus loin dans des comparaisons, sans une étude de définition technique beaucoup plus précise, qui dépasse largement le cadre de ce rapport, et nécessite l'intervention d'un cabinet-conseil.

Différents types de scénarios possibles

(nous nous appuyons ici sur le document détaillé Direction des Archives de France, sous-groupe « Architectures » COPIL, 30 novembre 2005).

1. *Scénarios de « front-office »*
   - 1.1. une interface unique de recherche et de consultation, permettant l'accès à des documents dans un format homogène (« visualiseur unique ») ;
   - 1.2. une interface unique de recherche mais des environnements de consultation multiples, adaptés à divers types de documents (« interface fédératrice »).

   Le choix pour une BNUE, suite aux orientations du comité de pilotage, s'est porté en faveur du scénario 1.1 (voir fiche), particulièrement adapté à des documents imprimés, en rapport avec les choix faits par les grands acteurs de l'Internet. Il a en effet été estimé que les environnements de consultation multiples (scénario 1.2) n'étaient pas conformes au souhait du Comité de Pilotage d'une identité et interface uniques de la BNUE pour l'internaute.

2. *Scénarios d'architecture technique ;*
   - 2.1. Architecture centralisée, de type base de données unique à la BNUE.
   - 2.2. Architecture médiane, de type Webservices, s'appuyant aussi sur OAI-PMH.
   - 2.3. Architecture décentralisée, entièrement basée sur OAI-PMH.

Nota sur les métadonnées et le protocole OAI-PMH

Revenons au concept de métadonnées, qui se trouve pas si fortuitement que cela à la croisée des deux métiers qui nous intéressent, le fort ancien métier de

documentaliste, et le métier de moteur de recherche (qui donne accès aux contenus Internet) :

- Le documentaliste produit de tout temps, du tambour Rodolex aux notices de catalogue actuellement sur Internet, des métadonnées, éléments descriptifs d'une œuvre : auteur, titre, éditeur, année, sujet,…
- L'essor qu'ont eu les moteurs de recherche sur Internet à leur création est dû au même travail de description par métadonnées qu'ont fait les créateurs de sites Internet : description de leurs pages par « métatags ».

Même si maintenant, les méthodes de repérage d'un site Internet ont sensiblement évolué, ce point de collusion initial méritait d'être signalé[4].

Le protocole OAI-PMH (Open Archive Initiative- Protocol for Metadata Harvesting, première version début 2001)[5] est susceptible de jouer ce même rôle de point de rencontre entre les Bibliothèques mettant leur contenu sur Internet et les moteurs de recherche. Quelle que soit l'architecture technique et organisationnelle de la BNUE, il jouera un rôle important à ce titre.

Ce protocole aide à mettre en place des sites et portails d'accès à des ressources documentaires. Il définit deux types d'acteurs :

- Les fournisseurs de données, qui exposent des métadonnées de manière passive en implémentant le protocole OAI ; ce sont des « entrepôts OAI ».
- Les fournisseurs de services, comme pourrait l'être la BNUE, qui recueillent de manière active et régulière (« moissonnage ») ces métadonnées ; ce sont des « moissonneurs OAI ».

Le protocole OAI permet une acception très large du concept de métadonnées, pouvant aller jusqu'à une représentation complète du document. Ce point est important dans la discussion qui suit des divers scénarios techniques.

Par ailleurs, les entrepôts OAI n'allant pas –pour des raisons de coûts- mettre à disposition des moissonneurs des formats très variés, il importe que, quelle que soit l'architecture choisie, la BNUE sache appréhender le protocole OAI.

Signalons que le premier moissonneur OAI au monde, celui de l'Université du Michigan, appelé OAISter, est managé par une française, Muriel Foulonneau, ancienne de la Mission Recherche et Technologie du Ministère de la Culture.

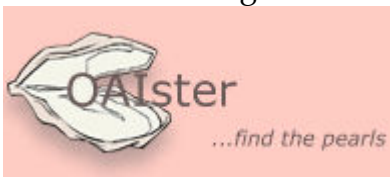


Figure 5.1 : OAIster, « find the pearls »
6 millions de notices documentaires moissonnées dans 575 institutions dans le monde

[4] Dans le même ordre d'idées, une collusion sémantique rarement soulignée concerne le mot « recherche » : recherche d'un document sur Internet à travers un moteur éponyme, mais aussi « travail de recherche » nécessitant documentation et bibliographie.

[5] Voir note de F. Nawrocki, Direction du Livre et de la Lecture, http://www.culture.gouv.fr/culture/dll/OAI-PMH.htm

http://oaister.umdl.umich.edu/o/oaister/viewcolls.html

OAIster est donné ici à titre de précurseur et de plus gros moissonneur OAI, sachant aussi que c'est un « moissonneur moissonné », à savoir que depuis mars 2004, les grands moteurs de recherche indexent les contenus OAI de OAIster, ce qui a multiplié par cent le trafic sur le site…

Discussion des scénarios d'architecture technique : scénario 2.1.

Le scénario 2.1 correspond au rôle joué par les moteurs de recherche sur les contenus Internet traditionnels (pages Web, documents) : ils vont « crawler » les sites et les copient dans leurs bases de données, au passage en les indexant, c'est à dire en créant une correspondance entre un mot donné et sa localisation dans une page donnée.
Tous les jours ou tous les deux jours, le moteur de recherche va vérifier si les pages répertoriées ont changé, ou si de nouvelles pages sont apparues (à titre d'exemple Google est supposé indexer 8 milliards de pages existant sur le Web).
Ceci correspond à un investissement informatique très lourd en termes de serveurs :

- Serveurs de bases de données, pour stocker les documents correspondants.
- Serveurs d'indexation créant les correspondances entre les mots et les pages les contenant.
- Puissance de calcul nécessaire pour l'indexation et l'affichage rapide des résultats de recherche.

Du point de vue du stockage de données, il n'est pas à exclure que dans ce cas, il y ait à récupérer le format-maître du document (le plus volumineux, un ouvrage = 20 Mo), car c'est sur celui-là que l'OCR est fait (pour avoir la meilleure qualité possible), donc c'est sur ce document-là, et pas sur des documents compressés, que la localisation du texte sur l'image (surbrillance) peut se faire. Cette récupération de documents-maître, si elle a lieu, devrait se faire de manière physique (CD-Roms) : compte tenu du volume de cette opération, elle ne peut avoir lieu en ligne.

Nous sommes là en plein dans le métier des grands moteurs de recherche, qui d'ailleurs ont bien compris qu'un contact le plus en amont possible avec les grands bibliothèques était nécessaire pour bien connaître les formats, et si possible réduire le coût de stockage des contenus (l'avantage d'indexation ira à celui qui aura numérisé les contenus, avantage de timing mais aussi avantage technique lié aux formats de numérisation et de stockage).

Si la BNUE devait se lancer d'emblée dans un tel scénario, une rapide projection de coûts , basée sur 10 fois le volume Gallica soit 1 millions d'ouvrages, donne une capacité de stockage nécessaire de 70 téraoctets, une bande passante de 1GB/s, un coût global d'investissement de 4M€, et un coût de fonctionnement annuel de 1,4M€ (10 personnes + maintenance + bande passante).

Discussion des scénarios d'architecture technique : scénario 2.3.

Bien que le scénario technique de type 2.3 totalement décentralisé ne soit pas compatible avec le scénario de « *front-office* » 1.1 (interface unique de consultation), nous donnons ici quelques exemples de sites de bibliothèques numériques mettant en œuvre ce scénario 2.3.

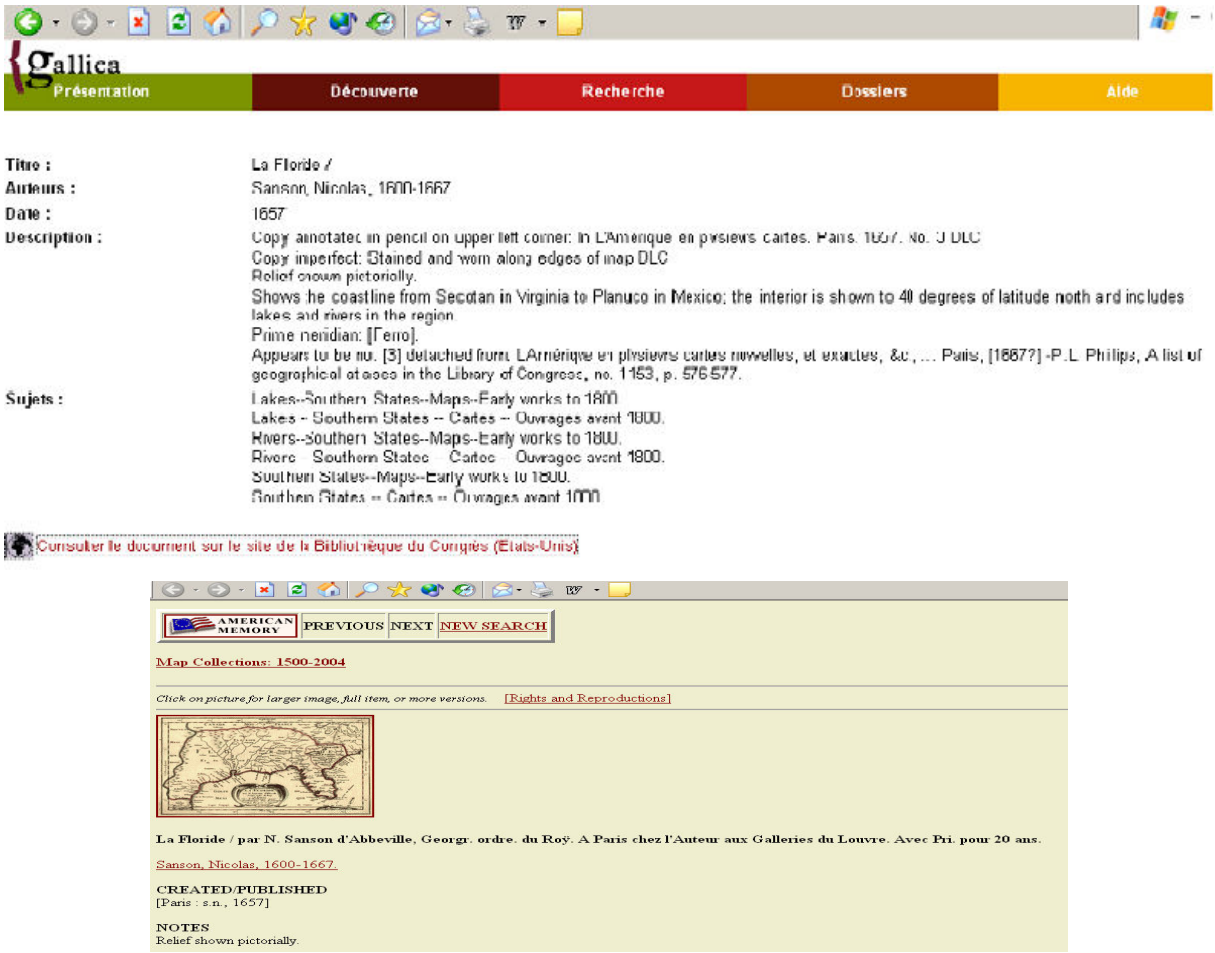


Figure 5.2 : exemple d'architecture décentralisée OAI, sans visualisation unique (scénarios 2.3 et 1.2), coopération entre BnF et Library of Congress
Sur une recherche Gallica (ex. Floride), le document LoC apparaît parmi d'autres documents Gallica ; une notice de métadonnées est consultable grâce à OAI ; puis le document est consultable en environnement LoC.

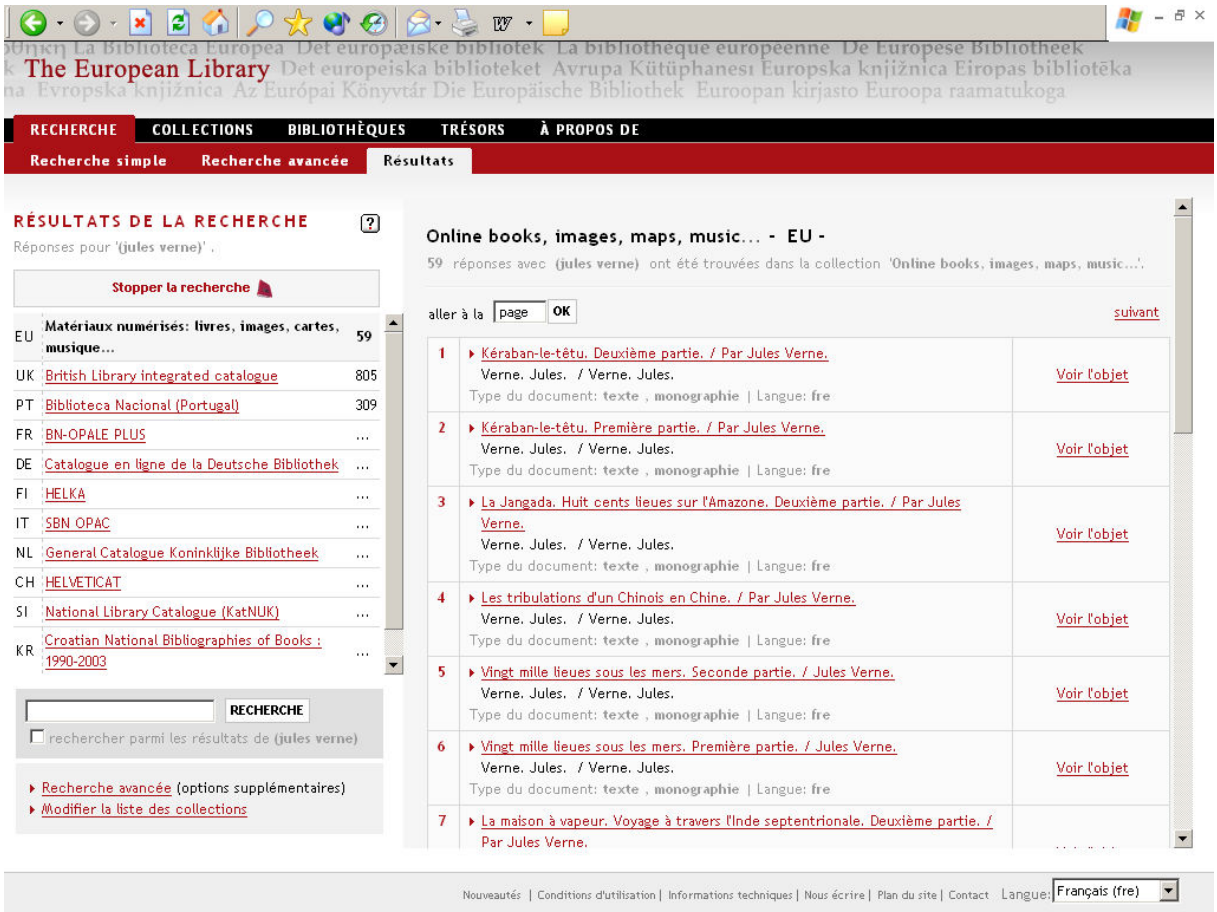


Figure 5.3 : exemple d'architecture décentralisée OAI, sans visualisation unique (scénarios 2.3 et 1.2), portail TEL « The European Library »
Sur une recherche dans TEL (ex. Jules Verne), on trouve 59 documents numériques sur Internet (et 1100 notices ce qui n'est pas d'un grand intérêt) ; à droite « voir l'objet », une fenêtre s'ouvre avec le visualiseur Gallica sur la première page du document.

Regardons l'aspect financier de l'entrepôt OAI dans le scénario 2.3, pour les deux raisons suivantes :

- ces entrepôts sont à présent une donnée du paysage (y compris à la BnF où nous avons vu, voir fiche, que 25 000 notices en OAI existaient déjà et étaient moissonnées[6]).
- ces entrepôts/fournisseurs de contenu ne vont pas investir (ni en argent, ni en temps de leurs développeurs informatiques) dans des formats d'interface spécifiques avec la BNUE, qui devra s'adapter à une architecture existante.

Un premier exemple économique est donné par OAIster (University of Michigan), qui fournit un coût d'investissement (hors salaires) de 45 000$ pour son serveur OAI. Un deuxième exemple est donné par la Direction des Archives de France pour un coût total de 60 000€ (y compris 20 000€ de salaires) de son serveur OAI.

[6] On pourra le vérifier en faisant une recherche d'auteur français (plutôt scientifique) sur http://oaister.umdl.umich.edu/cgi/b/bib/bib-idx?c=oaister;page=simple

Discussion des scénarios d'architecture technique : scénario 2.2

La mise en place d'un protocole de type « Web Services » permet de gérer les consultations en temps réel de serveurs distants ; le visualiseur BNUE affichant à la volée le document-source provenant du fournisseur de documents (note BnF/DSR/EMB du 28 novembre).

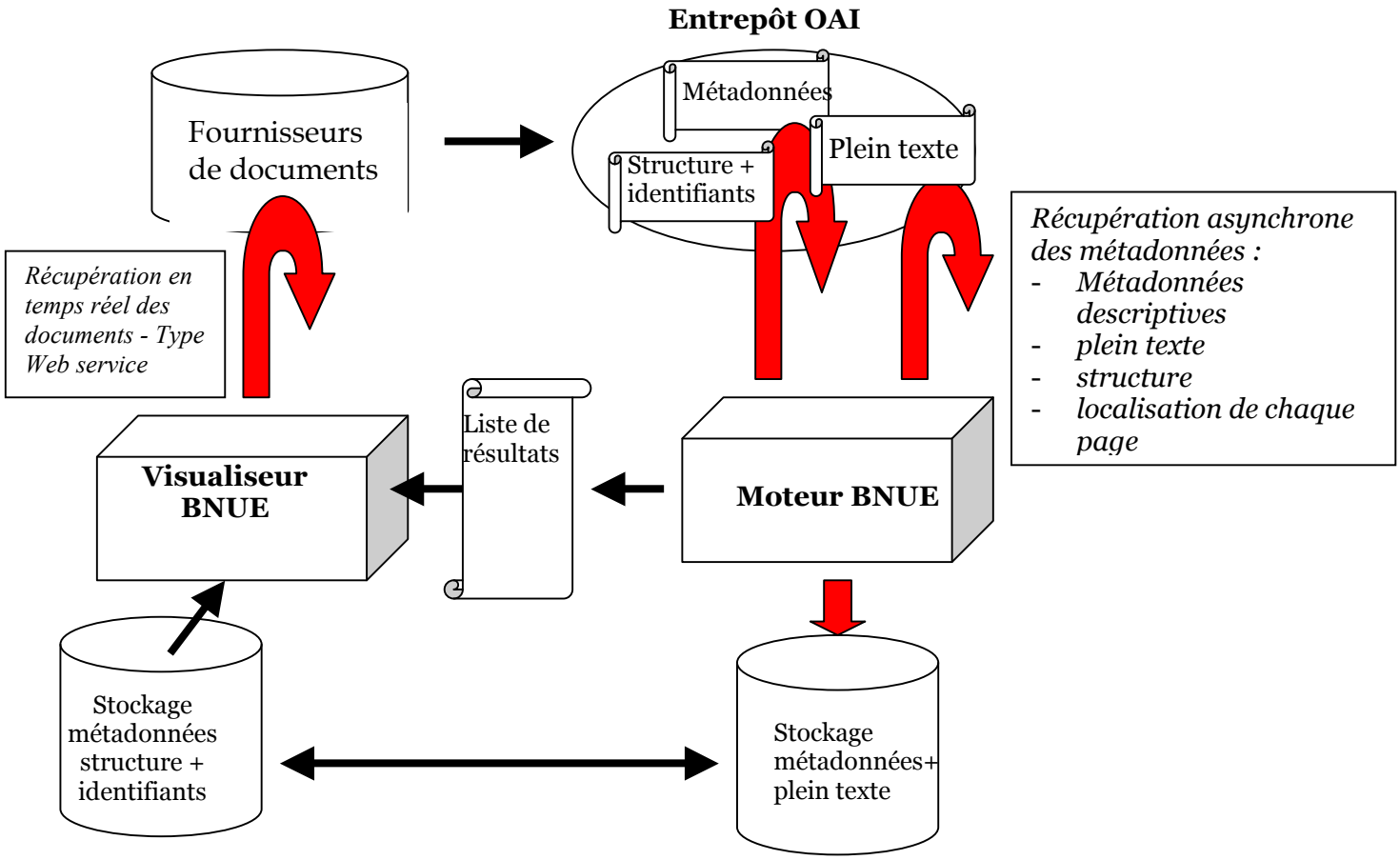


Figure 5.4, schéma issu de la note E. Bermès (BnF) du 28 novembre 2005

Il faut noter que, dans ce scénario comme dans le scénario décentralisé, le temps de réponse dépendra de la disponibilité du serveur de contenu du partenaire. Ceci n'est toutefois absolument pas différent des résultats ramenés par un moteur de recherche grand public, où la page est appelée sur le serveur de l'hébergeur du site, et où les messages « *serveur indisponible* » ou « *lien brisé* » peuvent apparaître – somme toute très rarement : dans ce cas-là l'internaute va voir la « copie en cache » de la page faite par son moteur de recherche sur ses propres serveurs (scénario 3.1).

Le scénario 2.2 est un scénario semi-intégré sur lequel nous pensons que la BNUE pourrait se baser dans un premier temps. Il concilie en effet les avantages du scénario 1.1 (visualisation unique BNUE), et les avantages économiques de l'OAI, aussi bien à la BNUE (pas de réplication des documents sur les serveurs BNUE dans un premier

temps) que chez le partenaire de contenu (coût raisonnable d'un serveur OAI, par ailleurs adapté à d'autres types de moissonneurs que la BNUE).

Dans un second temps, on ne peut exclure qu'il devienne nécessaire de copier les contenus sur les serveurs BNUE, et donc de passer au scénario 2.1 : cela signifiera que la BNUE aura un certain succès, ce qui pose le sujet différemment qu'il ne l'est *ab initio.*

> *Ceci pose, à terme, la problématique de l'identité BNUE par rapport à l'identité des propriétaires de contenu, sujet traditionnel d'Internet .*
>
> *Il est clair, d'ailleurs, que cette problématique est présente dans la stratégie « Book Search » des grands moteurs, puisque Yahoo présentait devant le Groupe Technologies (GT4) du comité de pilotage BNUE un transparent où figurait la phrase «* Attribution required for mass re-hosting *», ce qui semble signifier que le moteur Yahoo devra clairement faire apparaître l'origine des contenus qu'il recopie en masse.*

@@@@@@@

En conclusion, nous pensons que le scénario 2.2, qui constitue une innovation dans la présentation de documents sur Internet, **devra être validé progressivement, de manière expérimentale** : il est difficile de le décrire plus avant sur papier, ce n'est qu'en mettant en place des connecteurs WebServices avec des fournisseurs de contenus que l'on confirmera la validité opérationnelle et économique du scénario.

Dès que la BNUE aura commencé à mettre en œuvre de manière opérationnelle le scénario 2.2, elle pourra procéder à une étude de coûts sur la généralisation de ce scénario. Cette étude économique, à confier à une société spécialisée, devra décrire *a minima* les points suivants :

- Coûts de généralisation du scénario 2.2 à la BNUE ; par généralisation on entend passer au-delà d'un stade d'une dizaine de connecteurs entre la BNUE et des fournisseurs de contenus privilégiés.
- Coûts induits chez les fournisseurs de contenus pour se connecter à la BNUE.
- Coûts de migration de la BNUE vers les scénario 2.1 en cas de succès .

Nous proposons donc de commencer avec le scénario 2.2, en le validant de manière expérimentale et progressive, **en créant des « connecteurs BNUE » avec des fournisseurs de contenu intéressés.**

Fiche 6 : Schéma organisationnel BNUE

Cette fiche vise à projeter quelle pourrait être la structure fonctionnelle d'une BNUE au niveau européen.

Le sous-groupe « Architectures » du GT4 (Technologies) du Comité de pilotage a énuméré plusieurs scénarios d'organisations possibles pour la BNUE :

1. prise en charge par la BNUE de la recherche, de la consultation, de l'hébergement des données et de leur numérisation (*« organisation centralisée »)* ;
2. prise en charge par la BNUE de la recherche, de la consultation, de l'hébergement des données ; numérisation prise en charge par les contributeurs (*« organisation quasi-centralisée »)* ;
3. prise en charge par la BNUE de la recherche, de la consultation ; hébergement et numérisation pris en charge par les contributeurs *(« organisation distribuée »*) ;
4. prise en charge par la BNUE de la recherche ; consultation, hébergement et numérisation pris en charge par les contributeurs *(« organisation contributive ou décentralisée »*).

Suivant le Comité de pilotage du 17 octobre, l'organisation de type 4 n'a pas été étudiée plus avant, puisqu'il était souhaité une identité propre BNUE et une fonction de consultation dans un cadre BNUE.

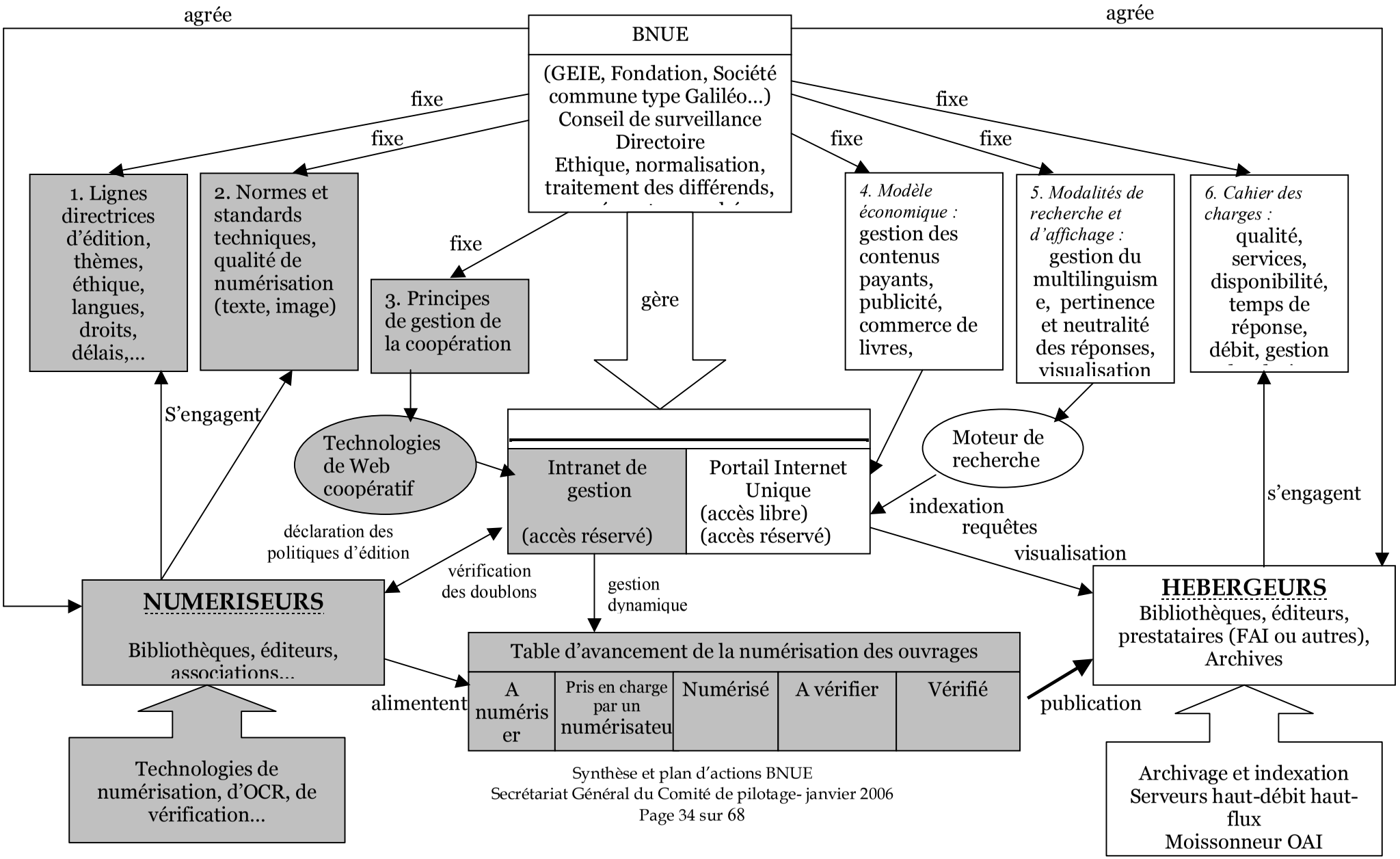
Le scénario décrit dans cette fiche se situe entre les scénarios 2 et 3 : la BNUE assure les fonctionnalités de recherche et de consultation, mais une nouvelle fonction « d'hébergeur » est introduite (sachant que cette fonction pourra être tenue par certains grands contributeurs comme les Bibliothèques nationales, mais sachant aussi que cette fonction pourra être prise en charge par des tiers du privé).

Il convient de souligner que ce scénario correspond à un schéma idéal de fonctionnement d'une telle bibliothèque au niveau européen : **il ne serait donc réalisé que par étapes**, en fonction de l'intérêt suscité auprès du public par la BNUE, et avec une forte volonté politique européenne pour la création d'une structure européenne quasi-centralisée.

@@@@@@@

Ce scénario serait bâti autour d'un portail internet unique, avec deux plans d'action différents :

- Les fonctionnalités de « back-office », propres à la numérisation d'un document, avant publication d'un document sur le portail BNUE : voir fonctions 1 à 3 en cadre grisé dans schéma ci-dessous.
- Les fonctionnalités de « front-office », propres à la consultation d'un document par l'internaute, après publication d'un document sur le portail BNUE : voir fonctions 4 à 6 en cadre non grisé dans schéma ci-dessous.


agrée
agrée
BNUE
(GEIE, Fondation, Société commune type Galiléo…)
Conseil de surveillance
Directoire
Ethique, normalisation, traitement des différends,
fixe
fixe
fixe
fixe
fixe
fixe
1. Lignes directrices d'édition, thèmes, éthique, langues, droits, délais,…
2. Normes et standards techniques, qualité de numérisation (texte, image)
3. Principes de gestion de la coopération
gère
4. Modèle économique : gestion des contenus payants, publicité, commerce de livres,
5. Modalités de recherche et d'affichage : gestion du multilinguisme, pertinence et neutralité des réponses, visualisation
6. Cahier des charges : qualité, services, disponibilité, temps de réponse, débit, gestion
S'engagent
Technologies de Web coopératif
Intranet de gestion
(accès réservé)
Portail Internet Unique
(accès libre)
(accès réservé)
Moteur de recherche
indexation
requêtes
visualisation
s'engagent
déclaration des politiques d'édition
vérification des doublons
gestion dynamique
NUMERISEURS
Bibliothèques, éditeurs, associations…
HEBERGEURS
Bibliothèques, éditeurs, prestataires (FAI ou autres), Archives
Table d'avancement de la numérisation des ouvrages
A numériser
Pris en charge par un numérisateu
Numérisé
A vérifier
Vérifié
alimentent
publication
Technologies de numérisation, d'OCR, de vérification…
Archivage et indexation
Serveurs haut-débit haut-flux
Moissonneur OAI

Description d'un back-office BNUE

Il est basé sur la relation entre la BNUE et les numériseurs « agréés ». Ce peuvent être des bibliothèques, des éditeurs, voire des associations[7]. Ils sont agréés suivant deux axes distincts :

- La conformité aux règles de numérisation fixés par la BNUE (axe technique).
- L'intérêt des contenus dans le cadre d'une ligne éditoriale fixée par la BNUE (axe scientifique).

C'est la BNUE en direct qui agrée suivant les normes qu'elle définit ; pour la conformité scientifique, elle se fait assister, dans chaque pays, d'un comité scientifique[8] qui donne son avis pour la validation scientifique, non par ouvrage, mais par programme de numérisation. Même si un numériseur est déjà agréé, un nouveau programme de numérisation qu'il propose doit recevoir l'agrément du comité scientifique BNUE pour y être intégré.

Vis-à-vis de ses partenaires professionnels pour l'approvisionnement en contenus (numériseurs), la BNUE remplit les fonctions suivantes :

- *(voir schéma, pavé grisé 1.)* Elle fixe une ligne éditoriale, et chaque année les numériseurs agréés déclarent auprès du comité scientifique de la BNUE une politique d'édition numérique conforme à cette ligne.
- *(voir schéma, pavé grisé 2.)* Elle fixe les normes et standards techniques de numérisation et d'affichage
- *(voir schéma, pavé grisé 3.)* Elle fixe les principes de travail coopératif : intranet de gestion, table d'avancement de la numérisation des ouvrages, modalités de vérification qu'un ouvrage n'est pas déjà numérisé ou en cours de numérisation

Description d'un front-office BNUE

Vis-à-vis de son public d'internautes-lecteurs, la BNUE garantit trois fonctions importantes :

- *(voir schéma, pavé clair 4.)* Elle établit les règles et les limites du modèle économique associé au portail : publicité par exemple réservée au monde du livre, gestion des droits et paiement pour les livres sous droits,…
- *(voir schéma, pavé clair 5.)* Elle assure le service de recherche de contenus en plein texte sur les contenus BNUE, voire par extension sur un certain nombre de contenus numérisés qu'elle indexe.
- *(voir schéma, pavé clair 6.)* Elle garantit, en liaison avec les hébergeurs, des règles de disponibilité des contenus et de temps de réponse aux requêtes.

Cette dernière fonctionnalité se fait via l'agrément d'hébergeurs de livres numériques. En effet le succès escompté d'une BNUE demandera, en régime de

[7] Ces deux dernières catégories peuvent le cas échéant avoir leur place en tant que numériseurs agrées par la BNUE, comme le montre l'état des lieux réalisé par le COPIL-GT1 (annexe au rapport); comme exemple d'association voir la bibliothèque d'articles de la Fondation Napoleon (http://www.napoleon.org/fr/salle_lecture/articles/index.asp) ; comme exemple de site de particulier numérisant des œuvres on trouvera http://maupassant.free.fr/pdf/horla.pdf)

[8] L'importance d'un conseil scientifique par pays est rappelée par J.N. Jeanneney dans son livre *Quand Google défie l'Europe*, page 96 : « Des conseils scientifiques….s'attacheraient au premier chef à l'héritage national. Ils seraient animés par telles ou telles hautes personnalités bénéficiant d'un grand rayonnement intellectuel et d'une autorité internationale ».

croisière, de faire appel à des professionnels de l'hébergement, capables d'assurer la gestion de serveurs à haut débit pour répondre aux requêtes des internautes.
Ces hébergeurs peuvent être de deux natures :

- soit des gros numériseurs agréés qui assurent leur propre hébergement, avec les règles requises par la BNUE ; mais on peut imaginer que ces numériseurs soient amenés à se concentrer sur leur cœur de métier qui n'est pas l'hébergement.
  Dans le cas de numériseurs agréés « de premier rang », ceci rejoint le schéma de l'organisation contributive, où « l'agrégateur institutionnel » peut être assimilé à cet hébergeur agréé, car il agrège les données des autres fournisseurs de contenu

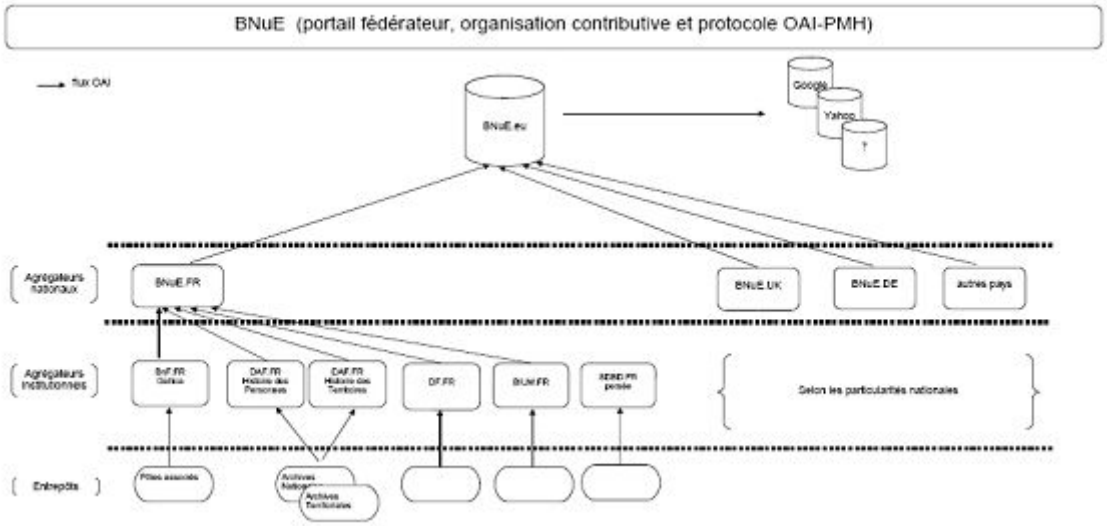


Figure 6.2 : extraite de l'étude Direction des Archives de France pour BNUE du 30 novembre 2005[9].

- conséquemment peuvent faire leur apparition d'autres types d'hébergeurs n'ayant rien à voir avec la numérisation de contenus : sociétés privées (hébergeurs ou FAI fournisseurs d'accès internet), gros serveurs universitaires (cf . OAIster, fiche 4)

Dans un cas comme dans l'autre, l'hébergeur peut être amené à jouer le rôle de « moissonneur OAI », permettant de récupérer les métadonnées descriptives des ouvrages des numériseurs agrées, leur fonction « entrepôt OAI » étant ouverte vers la BNUE uniquement dans le cadre du scénario d'architecture 3.2 (voir fiche 4).

Par ailleurs, dans le cas où se crée un marché de l'hébergement des livres numériques, des sociétés privées comme les FAI (qui bien souvent ont des filiales d'hébergement ou le sous-traitent) peuvent jouer ce rôle d'hébergeur ; parallèlement ils peuvent aussi jouer le rôle de « portail BNUE en marque blanche » (voir fiche) pour les internautes clients du FAI.

Enfin, il convient de signaler à titre subsidiaire (et ceci ne figure pas sur le schéma organisationnel de la page précédente) que dans un second temps les technologies de « Web coopératif » peuvent aussi jouer un rôle dans le « front-office » BNUE, i.e. les internautes peuvent être mis à contribution dans deux domaines :

[9] On fait figurer intentionnellement ces comparaisons, même si elles peuvent paraître difficiles à assimiler au lecteur externe aux groupes de travail COPIL, car elles permettent de constituer un langage commun entre membres de ces groupes, qui sera utile pour la suite du projet.

- signalement des corrections dans les erreurs de numérisation (Google Print le fait déjà, en mode signalement d'erreurs, voir annexe sur les formats de visualisation).
- sous modération BNUE, commentaires et forum de discussion sur les œuvres.

Fiche 7 : Présentation de la maquette COPIL du 11 janvier 2006

Les maquettes présentées au COPIL du 11 janvier, conformément aux décisions de celui du 17 octobre, visent à illustrer l'interface utilisateur, le format de visualisation unifié, la recherche plein texte.

Elles ont été réalisées gracieusement par partenariat avec deux sociétés que nous remercions:

- la société Isako, partenaire du site www.cairn.info, qui nous a été indiquée par un éditeur membre d'un groupe de travail du COPIL.
- La société Thomson, représentée au COPIL par son Président-directeur général.

Maquette Thomson, www.bnue.org

Cette maquette vise à illustrer ce que pourrait être une interface utilisateur de la BNUE. Du point de vue des contenus, elle vise principalement à donner un exemple de variétés de contenus, sous format de visualisation unique, et sous interface de recherche commune.

Comme précisé, elle a été bâtie sans architecture particulière (les temps d'affichage peuvent s'avérer un peu longs) et sans moteur de recherche spécifiquement élaboré, car ce n'était pas l'objet de la maquette.

Les contenus visualisables sont parmi les suivants :

- Une vingtaine de contenus Gallica mode texte retraités en mode image et texte (15 livres Balzac et 5 livres Châteaubriand).
- Un chapitre de livre prêté par un éditeur (Odile Jacob).
- Des contenus mode texte-image provenant de sites partenaires possibles : Cairn (portail de revues SHS des éditeurs), Numdam (Université de Grenoble- Société de Mathématique de France).
- Des contenus image uniquement, en format de visualisation unifié avec les autres contenus, rentrant dans la catégorie « documents images de grande qualité » (fiche 4, dernier paragraphe) : p.e. dans la maquette un livre manuscrit de Champollion, et un livre d'Ambroise Paré.

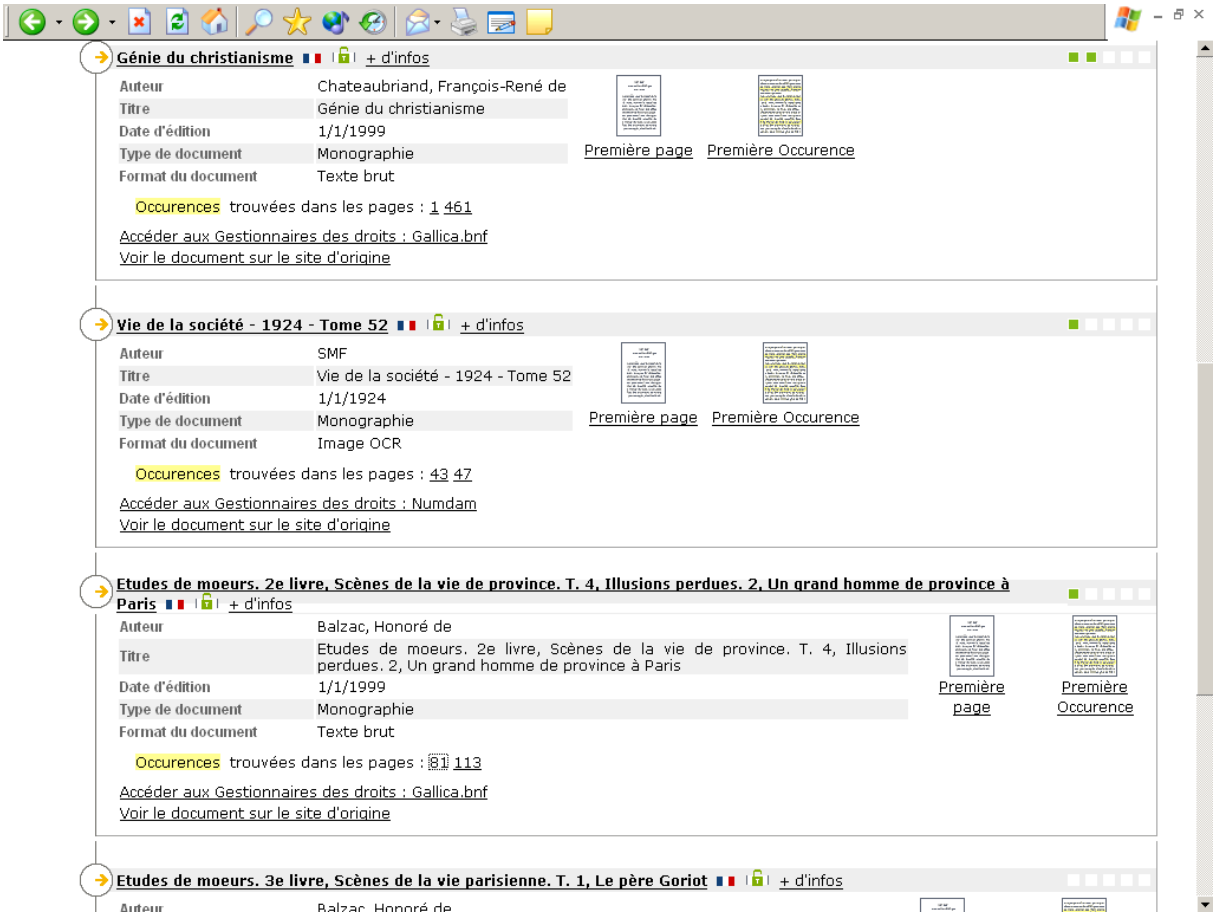


Figure 7.1, maquette ; une recherche Châteaubriand donne les œuvres de l'auteur (cf. « Génie du christianisme »), et des occurrences dans Balzac (« Illusions perdues »), ainsi que dans le bulletin de la Société de Mathématique de France en 1924 (site Numdam).

Maquette Isako, www.bibnum.org

Cette maquette a des contenus en nombre limité, et vise à montrer principalement l'aspect « traitement d'images » : OCRisation de contenus, retraitement qualité des images (marges, inclinaison), mise en mode image de contenus texte, traitement de fichiers images provenant d'un éditeur.

- Deux exemples d'OCRisation de livre Gallica qui avait été numérisé de manière directe (à partir du livre lui-même) : *Poincaré H., La science et l'hypothèse ; E. Poe, Histoires extraordinaires.*
- Deux exemples d'OCRisation de livre Gallica qui avaient été numérisés de manière indirecte (l'un à partir d'une microfiche, l'autre à partir d'un microfilm) : *Descartes, Discours de la méthode (1824) ; A. Comte, Cours de philosophie positive (1830).*
- Un exemple de contenu Gallica numérisé en mode texte uniquement (Acamedia, voir fiche 3 figure 3.3) : *Balzac, Le Père Goriot.*
  Pour ce contenu, où le mode image n'existait pas, il est reconstitué dans le format de visualisation du site-maquette, justifié, de lecture beaucoup plus agréable.
- Un exemple de contenu éditeur (chapitre d'un livre), prêté par les Editions Odile Jacob : *Hervé Zwirn, Les limites de la connaissance (2000).*

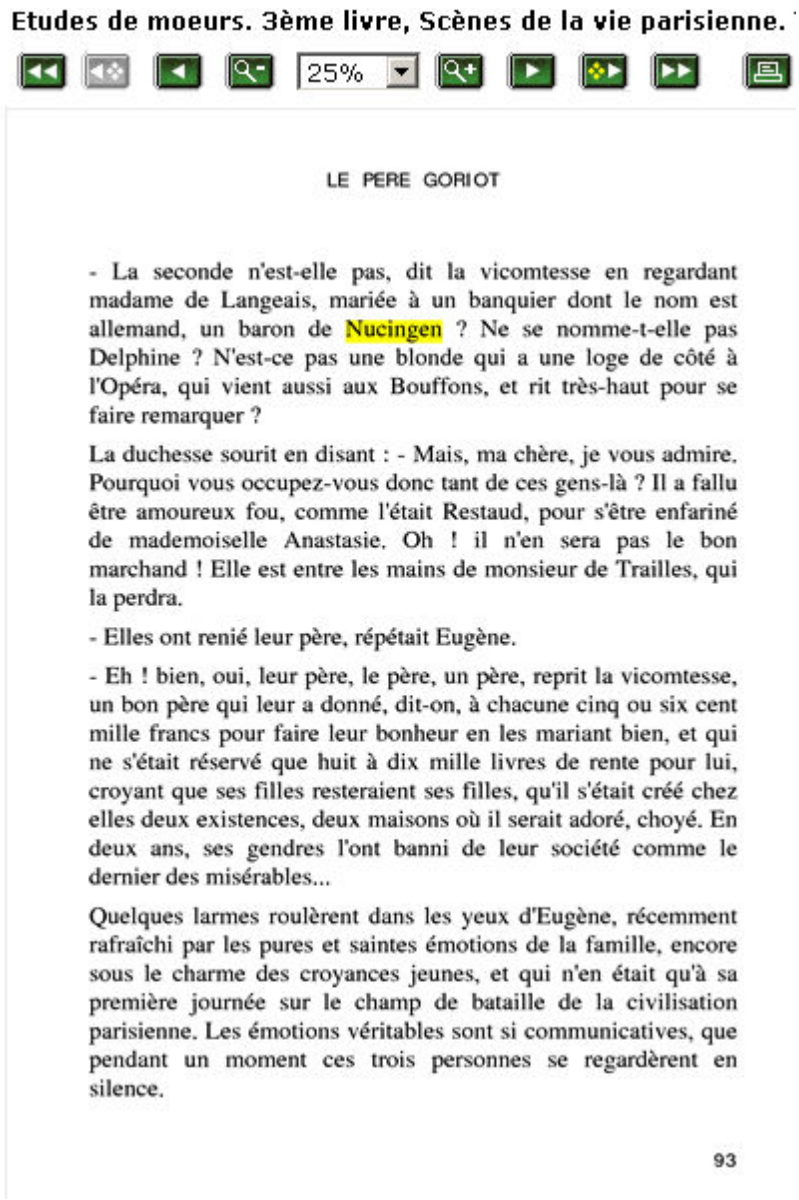


Figure 7.2 : maquette, visualisation en mode texte-image combinés, avec recherche plein texte, d'un contenu Gallica en mode texte uniquement. Visualisation plus conviviale y compris dans la pagination.

Fiche 8 : OCRisation des contenus Gallica

2006 : OCRisation des contenus Gallica

Compte tenu de ce qui précède (fiche 1 sur les modes de visualisation texte-image combinés des projets Google, MSN et de certains projets français, fiche 4 sur les modes envisagés de visualisation BNUE), le passage en mode texte-image combiné, par voie d'OCRisation, est une priorité pour les contenus Gallica en 2006.

Ceci signifie au passage, si la logique de standardisation des modes de visualisation pour le grand public et la conformité à la stratégie des grands acteurs de l'Internet sont suivies, qu'à **partir de 2006, toutes les numérisations Gallica devraient se faire ainsi**, en tout cas pour les documents OCRisables, disons postérieurs à 1800. Ceci signifie aussi que les budgets 2006 de la BnF consacrés à la numérisation (soit 2,5M€) devraient être consacrés dès que possible à l'OCRisation des contenus déjà numérisés.

Une projection des coûts correspondants sur 2006 se fait avec les hypothèses suivantes :

- ✓ on part des documents déjà numérisés (image) Gallica.
- ✓ on reste a priori dans un domaine postérieur à 1800 (60% de Gallica), les OCR sur documents plus anciens n'étant pas très fiables.
- ✓ C'est un OCR de qualité de base, taux de 95%, sans contrôle a posteriori. Ce taux est largement satisfaisant en première approximation.
- ✓ C'est un OCR avec combinaison avec le fichier image d'origine, permettant l'affichage en surbrillance, cf. fiche [ ].
- ✓ C'est un OCR avec image retraitée (recadrage en milieu d'écran de l'image, redressement de l'image quand l'image numérisée est penchée,...), c'est à dire conforme à la qualité de visualisation souhaitée.

La projection[10] sur 60 000 documents Gallica à OCRiser en 2006 se fait à 0,20€ la page (sur une fourchette de 0,12 à 0,20€, nous retenons le haut de la fourchette, tenant compte d'une qualité d'image retraitée), soit pour une moyenne de 300 pages par ouvrage, un coût de :

$$0{,}20€ \times 60\,000 \times 300 = \mathbf{3,\ 60\ M€}\ \text{en 2006}$$

Combien d'ouvrages dans une bibliothèque numérique ?

L'objectif quantitatif de contenus à numériser pour une mise en ligne (et non par souci de préservation) mérite qu'on s'y arrête.

L'utilité de la mise en ligne d'ouvrages libres de droits qui ne sont jamais demandés en salle n'est pas évidente ; certes l'offre peut créer la demande, mais aller au-delà d'un certain seuil pourrait correspondre à une dilapidation de ressources qui seraient sans doute mieux placées pour permettre une meilleure connaissance de ces

[10] ces chiffres sont à prendre avec précaution puisque le marché numérisation et OCRisation de la presse en juillet 2005 à la BnF a donné des résultats variant de 0,05 euros la page à 3 euros la page.

ressources par le grand public (partenariats marketing, publicité,…), voire pour soutenir dans l'optique d'une mise en ligne la numérisation d'ouvrages sous droits par leurs ayants droits.
Toutefois, des discussions menées avec les personnes impliquées sur ce sujet, il s'avère qu'une bibliothèque numérique accessible en ligne de 300 à 400 000 ouvrages (imprimés hors presse) répondrait très largement aux demandes du public visé par la BNUE. Pour Gallica, ceci correspondrait à une multiplication par 5 par rapport au volume actuel.

> *Pour mémoire , les chiffres prévus pour la British Library dans son récent plan de numérisation patrimonial (100 000 documents) concernent d'autres documents que les seules monographies et incluent des manuscrits, des cartes…..*
>
> *Dans un ordre d'idées plus proche de la configuration d'une bibliothèque numérique, rappelons que « Champion électronique » http://www.honorechampion.com/ met à la disposition de sa clientèle environ 150 000 textes numérisés en mode texte essentiellement centrés sur une période couvrant le Moyen Age au milieu du XIXè siècle.*
>
> *Le service américain NetLibrary http://legacy.netlibrary.com met à disposition des bibliothèques et universités américaines environ 80 000 ouvrages.*

En essayant de convertir en nombre de pages avec l'objectif de respecter une politique documentaire, on obtiendrait les chiffres suivants :

- presse quotidienne nationale rétrospective : 3.5 M pages (au moins jusqu'à 1935, plus en fonction des accords avec les journaux)
- monographies et revues : 400 000 volumes grand maximum, incluant les 80 000 volumes actuels, soit environ 120 M pages
- Total actuel Gallica: 25 M pages. Total à moyen terme : 125 M pages, hors images.

Il convient aussi de considérer, que pour un rythme actuel de numérisation de 5 000 ouvrages/an par la BnF, c'est un changement complet de process auquel il faudrait procéder, ne serait-ce que pour l'OCRisation en 2006 de 60 000 ouvrages.

En 2006, une étude et un test pour une numérisation de masse.

Conformément aux décisions du Comité de pilotage du 17 octobre, la BnF a passé un appel d'offres public[11] le 1er décembre pour une étude de numérisation de masse (y.c. reconnaissance de caractères) sur des machines à haute vitesse. L'étude porte sur la possibilité de numériser entre 800 000 et 1 500 000 ouvrages sur une période de 5 à 10 ans, soit environ 150 000 ouvrages par an. L'étude comportera en elle-même un test portant sur un échantillon représentatif, évalué par la BnF à 300 ouvrages.

[11] On trouvera le texte de cet appel d'offres à https://marchespublics.bnf.fr/dmp/dmp.nsf/webDetailM/0EB319CDACEC8F15C12570C3005193A9?OpenDocument ; réponse avant le 30 décembre 2005.

Concernant les machines de numérisation à haute cadence, non destructrices, il existe deux fournisseurs, et le test devrait porter sur ces deux fournisseurs:

- ✓ La société américaine Kirtas, qui a fait en France une campagne de promotion importante en 2005.
- ✓ Le groupe franco-suisse i2s/Assy, producteur de machines à numériser haute cadence avec lesquelles la société Infotechnique a monté fin 2004 en Moselle le centre de numérisation du cadastre d'Alsace-Lorraine pour le compte du Ministère de la Justice.

Ces deux sociétés sont en rapport étroit avec le projet Google Book Search, avec la particularité que Google a pris l'option de travailler directement avec des machines à numériser, sans passer par des spécialistes de la chaîne de numérisation (comme Jouve ou Infotechnique). Il semblerait que ceci s'en ressente, à la fois sur la cadence et sur la qualité de numérisation.

Un essai fait par la BnF le 21 septembre lorsque la machine Kirtas ATP Bookscan 1200 était en France a donné des résultats mitigés, notamment en termes de manipulations nécessaires, et le pourcentage des fonds BnF ainsi numérisables est estimé à 10%.
Concernant l'autre machine (Assy 4-Digital Book), il est à signaler que son utilisation en Moselle se fait sur des documents de calibre à peu près constants et très similaires entre eux.
L'étude de numérisation permettra sans doute de dégager quelques lignes directrices plus précises sur la capacité éventuelle de ces deux machines à gérer des fonds hétérogènes.

Le caractère non-destructif de la numérisation sur robot tourne-pages est certes un argument intéressant : on n'est pas obligé de « massicoter » un livre - et donc d'en avoir ou d'en acquérir un double - pour le numériser. Cet argument est toutefois à relativiser fortement en termes économiques, puisque le budget annuel dit « d'antiquariat » (achat de doubles de livres assez anciens pour massicotage) à la BnF se monte à 30 à 40 000€/an, soit plus significativement si on le rapporte au budget annuel de numérisation, 1% de ce budget.

A partir de 2007, une numérisation de masse ?

La numérisation en mode texte (autrement dit l'OCRisation) d'une part significative des contenus Gallica, au moins 60%, apparaît comme un objectif réaliste à partir de 2006. Au-delà de 2006, la question de la numérisation de nouveaux contenus, et son rythme, est posée.

Tout ce qui précède montre que la numérisation de masse de contenus comme seule stratégie BNUE serait difficilement et réalisable et explicable ; si la numérisation de masse de nos contenus français est en effet un point à étudier et à tester courant 2006, il ne saurait constituer la seule stratégie de la BNUE, notamment en 2006.

Les enjeux rappelés ci-dessus d'une meilleure connaissance des ressources numériques francophones par le grand public (l'accès ou « l'accessibilité » à la culture) sont tout aussi importants, sinon plus, y compris du point de vue financier. L'arrivée des grands acteurs de l'Internet sur ce marché donne un nouvel éclairage sur les annonces du premier d'entre eux en décembre 2004 : ce qui est important dans

l'annonce Google faite à ce moment-là, ce n'est pas tellement la quantité des contenus (sujette à discussion), mais bien le fait que Google rend ces contenus accessibles.

A cet égard, l'alternative entre une numérisation de masse, sur fonds publics par un organisme public, sans garantie d'indexation, à comparer avec l'autre alternative ouverte, une numérisation raisonnée, en partenariat technique et financier avec des acteurs de l'Internet, avec de meilleurs garanties d'indexation et d'accessibilité, doit être sérieusement soupesée.

Fiche 9 : Un partenariat d'accès avec les éditeurs

Le projet BNUE a suscité un vif intérêt du côté des éditeurs, comme l'a montré la participation active du SNE et des éditeurs suggérés par lui aux différents groupes de travail et réunions tenus sous l'égide du Comité de pilotage (GT1 à 5 entre le 1er septembre et le 17 octobre 2005, Sous-groupe architecture + réunions bilatérales en novembre et décembre 2005). Cet intérêt a été confirmé lors du déjeuner annuel du SNE tenu en novembre en présence du Ministre de la Culture, où les professionnels ont souhaité jouer un rôle plus important dans le projet, et où il leur a été proposé de nommer un représentant au Secrétariat Général du Comité de Pilotage, à partir du début de la phase opérationnelle à mi-janvier.

Le projet BNUE est en effet une occasion d'initier, dans une démarche associant le public et le privé, et de manière coordonnée, visible et portée politiquement, la mise en ligne sur Internet de contenus sous droits, à discrétion de chaque maison d'édition et de manière rémunérée.

On peut même dire –et ce fut une constante dans tous les groupes de travail – que le projet n'a de sens que si la BNUE donne à la fois accès à des contenus patrimoniaux mis à disposition du public gratuitement par les bibliothèques et à des documents protégés par le droit d'auteur mis en ligne moyennant rémunération par les éditeurs. C'est d'ailleurs ce que semble confirmer la configuration de Google Print qui pour l'instant comprend surtout des contenus sous droits à 70%.

Par ailleurs, une action commune entre pouvoirs publics et éditeurs sur la mise en ligne payante de livres aurait valeur d'exemple vis-à-vis des internautes, permettant de démarrer dès maintenant, à l'inverse d'autres secteurs culturels, une action pédagogique de fond et portée à la fois par les secteurs public et privé.

Côté BNUE, si dans un premier temps (2006), la proportion sera déséquilibrée en faveur des contenus patrimoniaux (la proportion de 80-20 a été mentionnée dans les groupes de travail), il est dans l'intérêt du public et donc de la BNUE de proposer de plus en plus de documents contemporains afin de répondre aux attentes des usagers (une proportion 50-50 pourrait être la cible au bout de deux ans).

Une plateforme mutualisée par la BNUE pour la rémunération des contenus sous droits

La BNUE doit être en mesure de proposer aux maisons d'édition les conditions-clefs de réussite suivantes :

1. accès unique aux ressources (interface de recherche unique et visualisation unifiée) ;
2. performance du moteur de recherche BNUE (transparence des critères de tri des résultats notamment) ;

3. solidité du système et performance des temps de réponse et d'affichage ;
4. visibilité du site BNUE vis-à-vis du grand public ;
5. solutions techniques adaptées et fiables pour le paiement.

Le canal BNUE ne doit avoir aucun caractère d'exclusivité pour les éditeurs ; tout au plus peut-on espérer, si les conditions-clefs de succès ci-dessus sont réunies, que la BNUE sera un canal prioritaire pour les éditeurs qui ne souhaiteront pas développer eux-mêmes sur leur propre site un tel service, et un second canal privilégié pour les éditeurs qui développeront un service semblable sur leur propre site[12].

C'est sur ces cinq axes, et notamment sur le dernier, que doit porter la coopération avec les éditeurs dans la phase opérationnelle de la BNUE.

Concernant les systèmes de paiement, il nous a été signalé par un éditeur de presse au cours de nos entretiens la naissance d'un nouveau système français en mai 2005 pour satisfaire les besoins du micro-paiement.

Il s'agit de Internet Plus www.internetplus.fr, système mutualisé créé par la quasi-totalité des fournisseurs d'accès français (AOL, Club-Internet, Cegetel, Neuf Telecom, Tiscali, Wanadoo) et les associations GESTE, GFII et ACSEL (dont les deux premières sont proches des milieux de l'édition). Ce système, opérationnel depuis juin 2005, permet la facturation des internautes par les FAI avec reversement aux sites de presse le mois suivant.
C'est un système sûr et qui favorise le micro-paiement (15 à 20% de prélèvement par les FAI contre 50% par les sociétés de cartes de crédits).

Ce type de coopération pour le micro-paiement avec les fournisseurs d'accès serait cohérent avec d'autres types de coopération que la BNUE pourrait avoir avec eux (par exemple accès BNUE sous leur propre marque, voir fiche [ ]).

Calendrier de travail – éditeurs et représentants des éditeurs.

1. Un secrétariat général élargi, ou mieux une structure privé-public (de type fondation, fiche 10) devrait permettre d'avancer rapidement sur la mise en œuvre opérationnelle dans le cadre du site-maquette BNUE (fiche 11) et d'une plateforme de rémunération des contenus sous droits.

2. Parallèlement, il conviendra d'avancer avec les éditeurs qui en ont manifesté le souhait ou en ont accepté le principe sur la mise en ligne par la BNUE des contenus de leur choix, cette liste n'étant pas limitative :
   - Editions Odile Jacob.
   - Editions La Découverte (groupe Editis).
   - Editions De Boeck[13]

[12] Dans ce cas, le partenariat entre la BNUE et l'éditeur devra se faire en parfaite articulation avec le développement du propre site de visualisation de l'éditeur.
[13] Les Editions de Boeck sont les premiers fournisseurs de contenus francophones sous droits au service Google Book Search (300 livres, voir fiche 1). Ils sont prêts à apporter leurs contenus à la BNUE.

- Editions Larousse
- Autres.

Les quatre éditeurs ci-dessus correspondent bien à la notion de « bibliothèque des savoirs » qui pourra être dans un premier temps la stratégie de la BNUE.

Ces actions 1. et 2. doivent pouvoir être finalisées pour mi-2006, en fonction de la date de mise en place de la structure de concertation et d'action commune.
Du point de vue des contenus sous droits, il semble exister aussi un gisement dans la littérature de sciences humaines et sociales de référence, à savoir les œuvres non rééditées d'auteurs majeurs du XX° siècle (Sartre, Aron,…). Elles pourraient être négociées entre la BNUE et les éditeurs concernés sous des conditions particulières.

Quel modèle économique pour ce partenariat, et pour la BNUE ?

Dans un modèle de BNUE reposant à 80% sur des contenus patrimoniaux, les éventuels revenus tirés des 20% de livres sous droits ne sont pas significatifs, et cette première phase ne saurait être considérée comme reposant sur un modèle économique, puisqu'elle serait essentiellement basée sur un financement public ou de mécénat.

Au-delà, avec une répartition plus équilibrée entre contenus patrimoniaux et contenus sous droits, la question d'un modèle économique peut être étudiée.

Les revenus tirés d'un site BNUE peuvent être de plusieurs natures :

1. Liens sponsorisés (publicités) liés au domaine culturel ou au domaine du livre sur les pages de recherche.
2. Pourcentage à verser à la BNUE sur la vente de livres en ligne par les éditeurs aux internautes arrivant par le portail BNUE.
3. Vente en ligne de produits dérivés : impression/photocopies d'ouvrages patrimoniaux (service actuellement assuré par la BnF), vente en ligne d'illustrations haute résolution (si la BNUE est étendue dans un deuxième temps à l'image).

A titre indicatif, Google GBS se contente uniquement de la source 1, et ne prend pas de commission sur les ventes de livres (source 2). Le modèle global de Google est d'ailleurs basé sur la publicité sur ses pages de recherche, à présent très rémunératrice aux Etats-Unis (mais pas en Europe semble-t-il), et sur le « page ranking » : mais ceci est le modèle de Google globalement, pas celui de GBS qui est un service sans modèle économique, à fonds perdus, au mieux un investissement marketing pour le service Google général.

Il est difficile d'aller beaucoup plus avant dans la définition d'un modèle économique BNUE : il est proposé d'avancer en marchant dans le cadre du partenariat privé-public, avec les éditeurs et avec les partenaires du domaine de la culture sur Internet (voir fiche ci-dessous).

Fiche 10: Quelle structure de partenariat privé-public ?

Nous avons cherché quelle pouvait être la meilleure structure de portage associant le public (BnF principalement, et en second lieu les Ministères), et le privé (Editeurs principalement, et en second lieu les industriels) ; cette structure devra avoir une personnalité juridique pour :

- Mener une concertation et un travail opérationnel entre public et privé à partir de février 2006.
- Gérer le budget BNUE dégagé par le Ministère de la Culture en 2006 (400 000€), ainsi qu'accueillir des fonds privés, ceux des partenaires comme ceux des mécènes.
- Négocier avec des partenaires de contenu de haut niveau (fiche 11) pour enrichir rapidement les contenus du site BNUE.

La structure associative, souvent utilisée dans le domaine culturel (cf. Michaël, cf. pôle de compétitivité IMVN) ne paraît pas forcément la structure la mieux adaptée pour la gestion de fonds publics notamment.
Une structure de portage intéressante est la fondation de recherche, ou fondation de projet à capital consomptible (sur la durée de vie du projet), telle que l'ont voulue les Ministres de la Culture et de la Recherche depuis 2003 (ces deux entités correspondent d'ailleurs aux partenaires ministériels du projet BNUE).

Les fondations de recherche (en application de la loi du 1er août 2003)

Ce statut constitue une avancée dans les outils à disposition et est un point important du plan gouvernemental pour l'innovation.
La dernière liste de fondations a été donnée par le Ministre délégué à la Recherche, voir http://www.recherche.gouv.fr/discours/2005/cpfondations.htm (23 mai 2005).
Pour une entreprise, le don ouvre droit à une réduction d'impôt de 60 % de son montant dans la limite de 5 pour mille du chiffre d'affaires.
Ces fondations sont d'abord – et c'est là l'innovation majeure - des ***fondations de projet***, à caractère consomptible, et n'ont pas grand'chose à voir avec les fondations connues dans le domaine culturel, comme celles créées autour de la succession d'un artiste.
Parmi les modèles de statuts en ligne proposés par le Ministère de la Recherche (http://www.recherche.gouv.fr/fondation/), l'option suivante paraît appropriée :

- Fondation à Conseil de surveillance et Directoire (et non simple conseil d'administration). Le Comité de pilotage BNUE pourrait poursuivre naturellement sa tâche comme Conseil de Surveillance de la Fondation.
- Fondation à membres de droit représentants des Ministères de l'Intérieur, de la Culture, de l'Enseignement Supérieur (et non simple Commissaire du Gouvernement représentant l'Etat).
- Fondation à Conseil scientifique placé auprès du Conseil de surveillance.

Exemples de fondations sur la diffusion des savoirs

La diffusion des savoirs est un des axes retenus par l'Etat pour le soutien aux nouvelles Fondations de recherche, qui doivent porter sur « des thématiques prioritaires et pertinentes, par exemple : développement durable, sécurité, énergie, environnement, diffusion du savoir ».

Sur ce dernier axe de diffusion du savoir, deux fondations public-privé ont été aidées par l'Etat au cours du deuxième semestre 2005 :

- *La première est la Fondation C.genial pour la culture scientifique et technique. Elle bénéficie des fonds suivants (lettres d'engagement signées à l'appui): EADS pour 1M€ sur quatre ans de 2006 à 2009, Schlumberger sur 1M€ pour quatre ans de 2006 à 2009, Areva sur 0,25M€ pour quatre ans de 2006 à 2009. Un soutien public de 0,6M€ a été décidé en décembre 2005 (voir site fondation EADS http://www.fondation.eads.net/default.asp?contentID=528)*

- *La deuxième est la Fondation pour le développement et l'attractivité de la science (en partenariat avec l'Académie des Sciences).*

- *Une liste des autres Fondations (hors diffusion des savoirs) figure sur le site du Ministère :*

Fondation de recherche pour le développement durable et les relations internationales, décret du 23 décembre 2004
Fondation Thérèse et René Planiol pour l'étude du cerveau, décret du 2 février 2005
Fondation santé et radiofréquences, (Orange, Bouygues, Sagem, TDF,…), décret du 2 février 2005
Fondation cœur et artères (Bonduelle, Sanofi, Auchan, Crédit du Nord…), décret du 2 mars 2005
Fondation bâtiment énergie, décret du 25 mars 2005
Fondation de recherche pour l'aéronautique et l'espace, décret du 1er avril 2005
Fondation pour une culture de sécurité industrielle, décret du 18 avril 2005
Fondation Institut Europlace de Finance, décret du 18 mai 2005
Fondation Garches, décret du 18 mai 2005
Fondation Habitat sans effet de serre (2004)CSTB, Lafarge, Arcelor, EDF,...
Fondation Recherche en alimentation (INRA, Bongrain, PernodRicard, Bel)
Fondation sécurité routière (INRETS, Peugeot, Renault,…)
Fondation Utilisation raisonnée de la sécurité animale (L'Oréal, Pierre Fabre, LVMH,…)

Il est clair que, au vu de ces exemples de fondations financées par l'Etat et trouvant un large soutien financier auprès de grands groupes industriels, le projet BNUE serait tout aussi légitime à figurer dans une telle liste, notamment dans l'axe « diffusion des savoirs ».

Quels partenaires pour une fondation française BNUE ?

Une structure privé-public, association ou plutôt fondation, regrouperait l'Etat (Ministères et Etablissements publics) et divers partenaires privés.

Du côté public, les partenaires seraient :

- Ministère de la Culture et de la Communication.
- Ministère délégué à la Recherche et à l'Enseignement Supérieur.[14]
- Bibliothèque Nationale de France.

La dotation prévue par le MCC à hauteur de 400 000€ (Loi de Finances 2006) serait une base de départ pour l'ouverture du site BNUE.

Du côté privé, on peut penser aux partenaires suivants :

- Syndicats professionnels : SNE (Syndicat National de l'Edition), GFII (Groupement français de l'Industrie de l'information), GESTE (Groupement des Editeurs de services en ligne).
- Des éditeurs partenaires techniques ou de contenu de la BNUE : grands groupes d'éditions, éditeurs (cf. fiche précédente).
- Des industriels intéressés à l'avancement du projet, et potentiellement partenaires techniques du projet :
  - ✓ Grand groupes Thomson, Thalès,…
  - ✓ Fournisseurs d'accès Internet éventuels partenaires : Wanadoo, Tiscali, Neuf, Free…
  - ✓ Moteurs de recherche : Yahoo !, MSN, Exalead,…
  - ✓ Libraires en ligne, partenaires publicitaires éventuels : Amazon.fr, Fnac, Alapage groupe France-Télécom)
- Des sociétés privées ou fondations intéressées au mécénat pour la numérisation de corpus et la mise en ligne :
  - ✓ Autres fondations françaises ou internationales : Fondation de France, Fondation Mellon,…
  - ✓ Mécénat de sociétés industrielles ou financières privées (banques, sociétés industrielles,…voir p.e. ci-dessus les partenaires industriels de la Fondation C.genial).

Pour les trois premières catégories d'acteurs (syndicats, éditeurs, industriels), on peut penser à une cotisation annuelle de 10 à 20 000€ suivant la taille des entreprises. Pour la troisième catégorie d'acteurs (mécénat), ils seraient considérés comme « membres bienfaiteurs » avec des donations *ad libitum*.

Autres structures de portage possibles

La fondation de projet nous paraissant le meilleur modèle, nous mentionnons ci-dessous quelques autres modèles de structures privé-public qui ont été évoqués.

1. Le modèle associatif est certainement le plus souple et le plus rapide à mettre en œuvre. Il n'est toutefois pas le plus idoine compte tenu de l'ambition politique portée dans le projet BNUE.

[14] Rappelons l'importance de ce partenaire ministériel sur de nombreux points : les contenus existants intégrables à la BNUE, l'avance technologique en mode texte-image combinés des sites dans son orbité, enfin l'aide qu'il peut nous apporter dans la constitution de comités scientifiques ad'hoc pour les contenus BNUE.

2. Il existe une société anonyme regroupant un établissement public et certains éditeurs : il s'agit de la société de droit belge CAIRN, au capital de 8M€; l'actionnaire principal est l'éditeur belge De Boeck avec 1,75M€ soit 27%, les sociétés d'investissement belges Gesval (1,5M€ soit 19%) et Meusinvest (1M€ soit 12,5%). Elle édite le portail de revues en sciences humaines et sociales www.cairn.info, ouvert en novembre 2005. Elle a bénéficié du soutien du CNL (Centre national du Livre) à hauteur de 38000€ en 2005, sur convention ouverte, pour la numérisation rétrospective de revues.
   Depuis la prise de participation qu'a faite la BnF dans CAIRN en décembre 2005 (à hauteur de 150 k€), on pourrait considérer que ce serait un vecteur possible d'une coopération privé-public. Il faut toutefois noter que les contenus de CAIRN s'adressent à un public très érudit, et ne correspondent pas au public visé par la BNUE.

3. L'idée a été émise que la BNUE pourrait se rapprocher d'un type de structure privé-public récemment mis en œuvre, les « pôles de compétitivité ». Il est certain que le pôle de compétitivité Image et Vie Numérique (un des quinze pôle de compétitivité à vocation internationale), qui regroupe notamment trois partenaires intéressés au projet BNUE (Thalès, le GET groupe des écoles de télécommunications dépendant du Ministère de l'Industrie, l'éditeur éducatif Odile Jacob multimedia), constitué sous forme d' association depuis décembre 2005 (http://image-idf.blogspot.com/) , et doté d'un budget de 150M€, pourrait accueillir avec intérêt le projet BNUE, si nécessaire.

4. Enfin, la possibilité d'être hébergée par une fondation existante comme la Fondation EADS (si celle-ci accepte de participer au financement du projet) pourrait être étudiée.

Un partenariat privé-public complémentaire avec les fournisseurs d'accès à Internet

En plus du partenariat privé-public suggéré avec les éditeurs privés, le plus important, nous suggérons ici un possible partenariat complémentaire à mettre en œuvre par la BNUE, reprenant ici une suggestion faite en GT2 du Copil par le représentant du SNE.
Il a en effet été suggère la possibilité d'un partenariat marketing de la BNUE avec les fournisseurs d'accès Internet (FAI), à savoir en France : Wanadoo (groupe France-Télécom), Club-Internet (groupe Deutsche Telekom), Neuf Télécom (groupe Telecom Italia), Tiscali, Free, AOL…Il est à signaler que la plupart de ces fournisseurs d'accès sont présents dans toute l'Europe, sous une marque ou une autre, ce qui peut être un avantage pour la partie européenne du projet.
Il s'agit pour la BNUE de vendre son portail «en marque blanche » à ces FAI qui le souhaiteront, dans un partenariat où le FAI assume sur son site les fonctions suivantes :

- Il assure la relation fonctionnelle avec l'utilisateur.
- Il assume la représentation de l'offre de contenus, suivant ses propres choix visuels et techniques.

- Il valorise les contenus avec ses propres outils (publicités, contenus tiers, liens avec des distributeurs, etc.).
- Il fournit des services à valeur ajoutée pour l'utilisateur afin de le fidéliser et de lui faciliter l'usage de la BNUE (outils de gestion des connaissances de type forums, blogs, conseils, communautés, veilles automatiques, liens avec sa boite mail, enregistrements dans des dossiers locaux, etc.).

Charge au FAI de réaliser les développements pour intégrer l'offre et les services BNUE dans son portail : la BNUE réalisera un premier serveur Web de ce type, afin de fournir un noyau fonctionnel minimal. C'est l'objet du prototype faisant suite à la maquette (voir fiche 11).

Ceci est susceptible de résoudre en partie un sérieux sujet auquel est confronté la BNUE, à savoir comment rendre l'accès le plus large à ses contenus via des sites grand public :

- Les moteurs de recherche Google, MSN, Yahoo, dans leur course à la numérisation et à l'indexation de contenus, n'ont pas ce problème puisqu'en eux-mêmes ils sont des portails grand public.
- Les moteurs de recherche européens de type Exalead, Sinequa (pour mentionner les sociétés auditées par GT4 du Copil), n'ont pas cette audience grand public permettant de démultiplier les accès aux futurs contenus BNUE. Par ailleurs ils n'ont pas cette stratégie de course à l'indexation de contenus lancés par les moteurs américains dans leur stratégie grand public.
- La coopération avec des acteurs européens comme les FAI, même si elle ne résout pas le sujet de l'indexation, permet d'améliorer accès aux et visibilité des contenus BNUE.

On peut même considérer, si nous réussissons à positionner à l'égard de ces interlocuteurs le projet BNUE comme *le* site de visualisation et d'agrégation des contenus de qualité, francophones dans un premier temps, européens par la suite, nous ouvrons la course à des acteurs européens de l'Internet, presque aussi importants que les trois moteurs de recherche américains.

Fiche 11 : Quel portail BNUE à mi –2006 ? (poursuite de la maquette et ouverture au public)

La commande d'avoir à mi-2006 (fin juin ou début septembre) une BNUE identifiable sur Internet, avec un volume de contenus significatif, est proposée comme suit, en poursuite et extension de la maquette qui est réalisée pour la 4° réunion du comité de pilotage du 11 janvier 2006.

La poursuite de cette maquette est proposée suivant quatre axes :

- Mise en place d'une architecture permettant de moissonner les contenus, et de tenir la charge pour la consultation.
- Partenariat technique (mise en place de connecteurs) et marketing avec les fournisseurs de contenus, dont les éditeurs.
- Mise en place de services à l'utilisateur, soit sur le site BNUE lui-même (moteur de recherche utilisant sémantique et linguistique), soit par partenariat avec des fournisseurs d'accès Internet français ou européens (fiche 9).

Partenariats contenus en 2006

Il est proposé de constituer au cours du premier semestre 2006 un « noyau dur » de contenus visualisables au sein du site BNUE.
La future structure de pilotage, ou le comité de pilotage, seront bien sûrs amenés à se prononcer sur ces contenus, mais un premier échantillon peut déjà être donné :

- Collection « Gallica classique » http://gallica.bnf.fr/classique/
  Cette collection, riche de 1 000 volumes de littérature classique, correspond en priorité aux objectifs de la BNUE :

> ***Extrait du site Gallica classique***
>
> *Gallica "Classique" est une collection du site Gallica. Cette sélection permet un accès direct aux textes fondateurs de la littérature française ainsi qu'une appropriation rapide. Conçu tant pour les lycéens et les étudiants que pour les professeurs, les chercheurs ou les curieux, ce site propose des outils de recherche et de navigation (chronologies, listes de sites) ainsi qu'un mode d'emploi. Les documents y sont disponibles, comme dans Gallica, en mode image et/ou en mode texte.*

Les volumes de littérature concernés se décomposent ainsi :

1. 140 volumes mode texte de Balzac et Chateaubriand (partenariat avec l'éditeur numérique Acamedia, disparu), dont un figure dans la maquette COPIL du 11 janvier 2005.
2. 100 volumes mode texte de littérature « Classiques Garnier-Flammarion » (partenariat avec l'éditeur numérique Bibliopolis, disparu), très diversifiés.
3. Le reste soit 850 volumes en visualisation Gallica traditionnelle mode image.

Il est certain que ces 1 000 volumes de grande littérature française gagneraient à être les premiers visualisables en mode texte/image combinés au sein de la BNUE (1. et 2. sont en mode texte sans image, pour 3. c'est l'inverse).

- Au-delà, les contenus qui seront OCRisés en 2006 (fiche 8) suivant la politique éditoriale Gallica pourront être intégrés au fur et à mesure dans la BNUE.

- Intégration d'autres contenus, notamment ceux émanant des membres du Comité de pilotage :
  - ✓ Rapports Documentation française.
    Les rapports Documentation Française ne sont généralement pas en mode texte, mais ils peuvent faire partie des contenus mode image traités par le visualiseur BNUE, en proportion limitée (cf. fiche 4).
  - ✓ Revues SHS Persée (Direction de l'Enseignement Supérieur)
  - ✓ Revues SHS Portail éditeurs CAIRN (dont La Découverte groupe Editis)
    La totalité de ces contenus n'est pas forcément à intégrer, peut-être une sélection à voir avec chacun des éditeurs de sites concernés.

- Documents numérisés Maison de l'Orient Méditerranéen[15] (Université Lyon II comme Persée) http://www.mom.fr/bibliotheque/bibnum/
  - ✓ Une cinquantaine de documents d'archéologie en français, en allemand, en italien : c'est une petite BNUE qui peut intéresser les partenaires européens.
  - ✓ Les documents ne sont pas en mode texte (p.e. livre manuscrit de Champollion ci-dessous), mais peuvent être traités par le visualiseur BNUE.

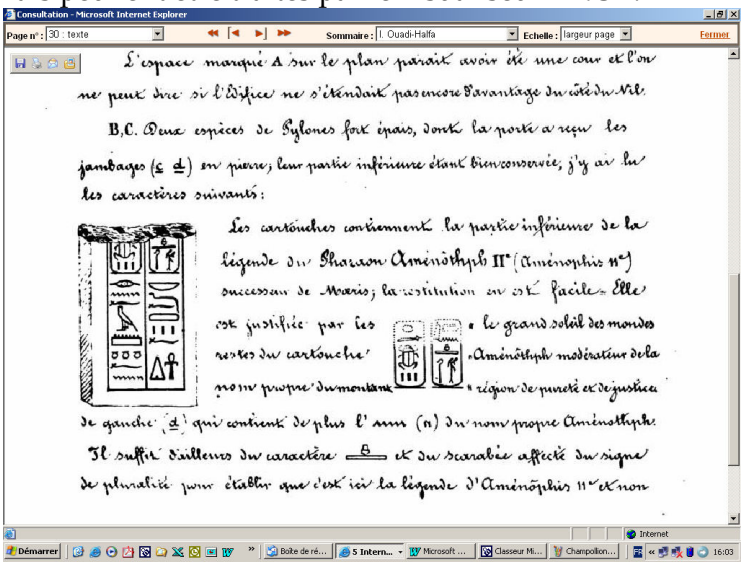


Figure 11.1, livre manuscrit de Champollion, site Maison de l'Orient et de la Méditerranée (MOM), visualisation MOM

Ce type de contenus pourrait être intégrés sous visualisation BNUE en accord avec MOM.

- Livres anciens complémentaires de Gallica :
  - ✓ Bibliothèques Virtuelles Humanistes (Centre d'Etudes Supérieures de la Renaissance, Université de Tours) http://www.bvh.univ-tours.fr/accueil.asp 87 ouvrages
  - ✓ Bibliothèque Interuniversitaire de Médecine (BIUM, Université Paris V) http://www.bium.univ-paris5.fr/histmed/medica.htm ; sur les 3200 ouvrages numérisés, intégrer en priorité ceux qui le sont en mode texte.

[15] Ces contenus numérisés sont signalés à juste titre par J.N. Jeanneney dans son livre « *Quand Google défie l'Europe* », pages 21-22.

- La bibliothèque de l'Université du Michigan contient 25 000 ouvrages numérisés en mode texte et image, dont on peut raisonnablement que 3 à 5 000 ouvrages en français (ou pas seulement en français, voir exemples ci-dessous) pourraient rejoindre la BNUE (http://www.hti.umich.edu/g/genpub/) :

Figure 11.2 : recherche « Poincaré » dans la base Université du Michigan, trois résultats « européens »
1. *(à droite) Henri Poincaré, Die neue Mechanik (La Mécanique nouvelle, **en allemand**)*
2. *(à gauche) Œuvres de H. Poincaré publiées après sa mort par G. Darboux (**en français**)*
3. *(au centre) Raymond Poincaré, The origins of the war (1922, traduction **en anglais** de Les Origines de la guerre 1921)*

(à signaler que 2. et 3. ne sont pas numérisés dans Gallica, et de toute façon il s'agit là de contenus mode texte/image)

(Ces trois livres 1. 2. 3. apparaissent dans les résultats de recherche Google Book, puisque UofM est un des partenaires privilégiés de Google, mais ne sont pas visualisables sur GBS alors qu'ils le sont sur UofM)

Les avantages d'une coopération immédiate de BNUE avec l'Université du Michigan, sur les livres déjà numérisés par eux, méritent qu'on s'y attarde :

- ✓ il s'agit de la plus grande base numérisée de livres texte/image au monde (25 000 ouvrages), plus importante que Google Book Search.
- ✓ Elle contient un nombre d'ouvrages francophones, ou d'ouvrages français traduits dans d'autres langues, tout à fait appréciable (cf. exemples ci-dessus). Ce n'est pas un hasard si Google est allé chercher ce partenaire-là en premier.
- ✓ D'un point de vue technique, UofM joue avec son site OAIster (voir fiche 4) un rôle d'entrepôt/moissonneur OAI dont nous pourrions nous inspirer pour la BNUE (coopération technique, et pas seulement en termes de contenus). Ce site OAIster est managé par une française ex-Ministère de la Culture.
- ✓ Le fait aura un impact que nous reconnaissions le travail de numérisation fait par UofM, et que nous montrions être prêts à travailler avec eux, comme eux travailleraient sans doute avec plaisir avec la BNUE.

Pour nos futurs contenus dans la BNUE (contenus francophones ou rayonnement de livres français traduits dans des langues européennes, cf. figure 11.2 ci-dessus), la coopération avec UofM pourrait être plus immédiatement rentable et opérationnelle qu'une coopération avec une Bibliothèque nationale européenne à ce stade, *y compris dans l'optique européenne de la BNUE* (cf. livre en allemand et en anglais figure ci-dessus).

- Le cas échéant, la bibliothèque numérique canadienne sans doute la plus utilisée en France (Université du Québec à Chicoutimi)
  - ✓ les classiques des sciences sociales et sciences sociales contemporaines.
  - ✓ http://www.uqac.uquebec.ca/zone30/Classiques_des_sciences_sociales/, voir collections 2 et 3.
  - ✓ 500 ouvrages hors droits, 500 ouvrages sous droits cédés à cette bibliothèque numérique (à vérifier pour ouvrages sous droits).

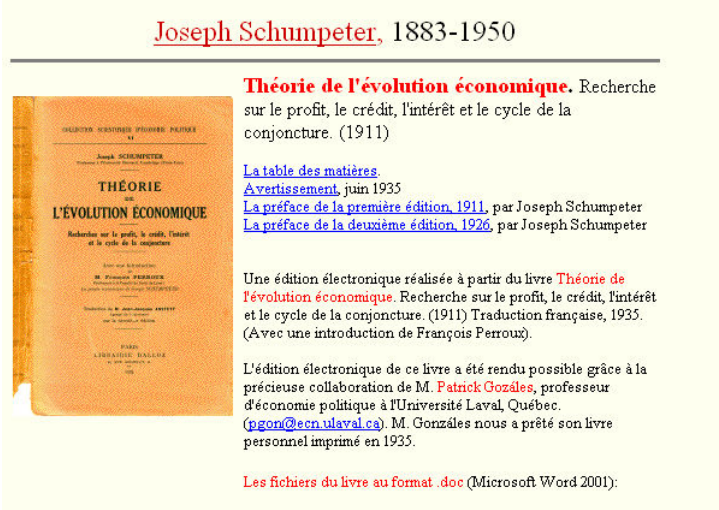


Figure 11.3, exemple d'un des 500 classiques sciences sociales Université du Quebec (dans cet exemple livre non numérisé Gallica)

Ce type de contenus pourrait être intégré sous visualisation BNUE en accord avec Université du Québec.

- et bien entendu, pour conclure, intégration de contenus sous droits (livres en totalité ou partiels) en coopération avec des éditeurs français volontaires, fiche 9).

Développements techniques – coût 2006 de la BNUE.

Bien qu'il soit difficile de donner un budget avant les orientations qui seront arrêtées par le COPIL du 11 janvier 2006, nous proposons ici un budget de fonctionnement raisonnable d'un site BNUE sur l'année 2006.
Ceci est valable pour un site BNUE francophone (en attente de décision européennes en la matière, 2006 est une année de transition), et hors budgets d'OCRisation et de numérisation (qui sont donnés par ailleurs, fiche 8).

Ressources	Emplois
➢ MCC dotation budgétaire.....400 000 ➢ Dotation des Fondateurs (éditeurs, industriels).....300 000 ➢ Mécénat.....300 000	➢ Personnel (3 p.).....225 000 ➢ Développements techniques : ✓ Connecteurs BNUE avec une quinzaine de sites..... .....150 000 ✓ Plate-forme d'accès et de services (dont moteur).....300 000 ✓ Gestion des droits et des paiements.....60 000 ✓ Adaptation à l'environnement FAI.....40 000 ➢ Etudes techniques et économiques : ✓ Etude technique et financière d'architecture.....50 000 ➢ Matériels informatiques (serveurs BNUE).....75 000 ➢ Promotion et marketing du site BNUE.....100 000
Total 1 000 000 €	**Total 1 000 000 €**

Ce projet de budget très indicatif appelle les commentaires suivants :

- Le nombre de personnes en année pleine est évalué à trois (animateur, responsable technique, responsable partenariats et marketing) ; il peut être réduit en fonction de mises à disposition éventuelles par des membres fondateurs de la Fondation.
- Les connecteurs BNUE avec une quinzaine de sites (cf. par exemple sites ci-dessus) visent à amorcer la pompe, pour avoir une masse de contenus significative à l'ouverture du site BNUE (juin ou septembre 2006).
- L'étude technique et financière BNUE est celle mentionnée en fiche 5, à lancer au premier trimestre 2006, visant à évaluer le coût d'architecture 3.2 en rythme de croisière, à partir de fin 2006, et un passage au scénario 3.3, en cas de succès immédiat de la BNUE.
- Les dépenses de promotion et de marketing sont absolument indispensables, notamment à l'ouverture du site, pour le faire connaître en dehors d'un public d'initiés.

Fiche 12 : Poursuite du portage BNUE au niveau européen

Il ne saurait y avoir de BNUE sans coopération européenne. Enclencher cette coopération est l'objectif principal, non atteint à ce jour (janvier 2006) ; diverses actions ont été entreprises pour atteindre ce premier objectif.
Le démarrage de cette coopération, qu'elle ait lieu sous l'égide de la Commission ou non, et les premiers résultats de cette coopération, demanderont sans aucun doute du temps.
D'un autre côté, il est clair que la course à la mise en ligne des contenus patrimoniaux (bibliothèques) et d'éditeurs (livres sous droits) est lancée au niveau mondial, avec déjà un certain nombre de résultats, pas toujours valorisants pour les contenus francophones (voir fiche 1). La France doit être à même de tenir au plus vite son rôle dans la défense et la mise en ligne de ce type de contenus francophones sur Internet, et c'est un second objectif tout aussi crucial, sinon plus.

C'est pourquoi il est proposé que l'action de portage de la BNUE au niveau européen soit ***une des quatre actions prônées, une brique parmi les différentes briques***, qui ne doit pas nous empêcher d'avancer dans la réalisation de ce second objectif.

Actions européennes entreprises sous l'égide du Comité de pilotage.

Il convient là de distinguer deux niveaux d'action différents, complémentaires et tout aussi importants l'un que l'autre :

1. L'action des Gouvernements auprès de la Commission Européenne.
2. La coopération entre bibliothèques nationales européennes sur le projet BNUE.

Concernant le premier point, la réponse actuelle de la Commission à la lettre de six chefs d'Etat du 28 avril 2005 est [16]:

- La communication de la Commission au Parlement en date du 30 septembre 2005 (court document de 14 pages), sur un plan « i2010 : Digital Libraries ».

> *Une analyse de ce document à la lumière de nos axes de travail retient les points suivants concernant l' action « i2010- Bibliothèques numériques »:*
> - *Elle concerne le patrimoine culturel, et non l'information scientifique et technique qui sera traitée séparément en 2006.*
> - *Elle porte sur tous les documents multimédia, pas seulement sur l'imprimé.*
> - *Elle indique le partenariat entre le public et le privé comme très souhaitable.*
> - *Elle rappelle le rôle important des bibliothèques nationales européennes sur le sujet.*

[16] On trouvera la page de la Commission et les différents documents mentionnés à http://europa.eu.int/information_society/activities/digital_libraries/index_en.htm.

- *Elle reste vague dans ses objectifs :* « une possible recommandation en 2006 pour la numérisation et la conservation numérique »
- *Elle rappelle la source de financement eContentplus, dotée de 60M€ sur 2005-2008 pour* « l'accessibilité et la facilité d'utilisation du contenu culturel et scientifique européen »

A contrario, elle ne mentionne rien sur une possible « Bibliothèque numérique européenne », et son titre au pluriel « Bibliothèque numériques » *en est un signe.*
A contrario aussi, elle ne mentionne aucun des deux projets qu'elle finance de manière importante et ayant trait au sujet : Michael et TEL (voir fiche 13).

- Une consultation en ligne réduite, de deux pages et huit questions, pour réponse avant le 20 janvier 2006 des « différentes parties intéressées ».
  Le Secrétariat général pour les Affaires européennes est en charge de la réponse du Gouvernement français à cette consultation ; deux réunions ont eu lieu en novembre et décembre ; la réponse, à laquelle nous avons été associés, reprend dans les grandes lignes et dans les limites du champ des questions posées certains axes développés dans le présent rapport.
  Toutefois, la forme-même de ces consultations fait que plusieurs contributions peuvent venir d'un même pays (p.e. en France réponse SGAE, réponse BnF, autres acteurs le cas échéant,…).
  Enfin, aucun délai n'est indiqué pour la synthèse et la restitution, qui restent dans l'objectif non précisé d'*« une possible recommandation en 2006 pour la numérisation et la conservation numérique ».*

Ces deux documents apparaissent clairement comme la réponse de la Commission à la lettre du 28 avril des six chefs d'Etat ; à ce stade, il paraît difficile d'infléchir la position de la Commission, en tout cas pas avant le dépouillement complet des réponses à sa consultation en ligne, c'est à dire au mieux pas avant juin 2006.

Concernant le point 2. ci-dessus, il a paru important, en application des décisions du Comité de pilotage du 17 octobre, qu'une démarche opérationnelle soit faite auprès des Bibliothèques nationales d'un certain nombre de pays européens. Dans cet objectif, une lettre précisant les premières pistes d'une BNUE que nous avons pu formuler (contenus possibles, architectures possibles, etc.) a été envoyée en date du 6 décembre 2005 par le président de la BnF aux responsables des bibliothèques nationales suivantes (voir mail du Pdt. Jeanneney en annexe). Cette lettre est la première proposition concrète de coopération avec ces Bibliothèques européennes.

Il convient toutefois de noter, en ce qui concerne la Commission européenne, que le bilan en termes de projets communs (cf. Bibliotheca Universalis projet éteint, cf. TEL cf. Michaël, voir fiche 13) dans le domaine des bibliothèques numériques n'est pas satisfaisant en termes de public visé, en termes de public atteint (ces sites ont un caractère très confidentiel), en termes de temps de déploiement des projets. C'est d'ailleurs sans doute consciente de ces semi-échecs que la Commission hésite à s'engager dans un nouveau projet comme la BNUE.

Une urgence pour les contenus francophones

Les actions 2 (fiches 9 & 10, partenariat public-privé) et 3 (fiche 11, poursuite du site) que nous proposons visent à avancer en 2006 dans l'élaboration d'un site de contenus francophones, pour que ce chantier avance.

Ce site peut aussi représenter une force dans les négociations qui seront menées, y compris au niveau européen, avec les grands acteurs de l'Internet sur le sujet (voir fiche 2 sur la stratégie des acteurs de l'Internet et voir ci-dessous).

La francophonie n'est certainement pas sur le même pied d'égalité que les autres langues européennes, le public qu'elle concerne sur Internet et sur d'autres supports est beaucoup plus important. A cet égard, au niveau des partenaires BNUE, seule la position de la Bibliothèque d'Espagne pourrait être comparée (ses programmes de numérisation démarrent).

Les coopérations immédiates proposées par exemple avec les contenus francophones déjà numérisés Université du Michigan, Université du Québec, peuvent s'intégrer par la suite dans une coopération européenne qui suivra son rythme.

Enfin, comme nous l'a suggéré un représentant du SNE, **le Salon du Livre consacré à la Francophonie en mars 2006** (et non à un pays spécifique comme la Chine ou l'Italie ces dernières années) peut être un espace idéal pour faire avancer médiatiquement le sujet d'un versant francophone et immédiat de la BNUE.

Une action vitale et de première nécessité : une réponse coordonnée des bibliothèques nationales européennes aux sollicitations des acteurs de l'Internet ?

Au rythme où se sont enchaînés les événements depuis l'installation du Comité de pilotage le 13 juillet 2005, on peut se demander si une action de première nécessité sur ce chantier 4 ne serait pas que les Bibliothèques nationales des pays d'Europe se mettent d'accord, sous l'égide de leurs gouvernements respectifs, pour se positionner ensemble et donner une réponse commune aux trois acteurs de l'Internet (Google, MSN, Yahoo) qui ont entamé la course à la mise en ligne et à l'indexation des contenus de qualité sur Internet, et qui les sollicitent en permanence depuis deux mois.
La différence – et elle est de taille – avec un projet BNUE *ex cathedra* dont on peut mettre un certain temps, en tout cas au niveau européen, à définir les contours, est que dans ce cas les pays européens se positionnent face à une démarche existante. Ceci peut forcer l'Europe à agir vite, et n'est pas forcément un abandon des objectifs européens de la BNUE décrits dans le présent rapport.

Un premier coup de boutoir (peut-être rattrapable) a été porté à cette stratégie par l'accord MSN - British Library, mais il est encore possible que les Bibliothèques nationales de tous les pays européens se mettent d'accord pour répondre de manière commune aux sollicitations des trois acteurs de l'Internet.

Cette réponse commune devait se faire sous forme d'une entité commune émanant des Bibliothèques nationales, cautionnée par les gouvernements européens, entité chargée de la négociation avec ces acteurs, et ayant plus de poids sur la négociation de sujets assez délicats (comme le « *page ranking* », par exemple) .
Même si le partenariat avec le privé (éditeurs pour les livres sous droits) doit être laissé provisoirement de côté dans cette option, ce qui constitue un inconvénient[17], l'avantage majeur en est qu'elle pourrait donner naissance effective à une BNUE plus rapidement et plus efficacement que par les canaux normaux.
Une répartition des rôles entre entité commune et Bibliothèques nationales pourrait être :

- A l'entité commune l'exclusivité des négociations avec les acteurs mondiaux de l'Internet ; la conclusion avec un ou deux d'entre eux aurait l'avantage de résoudre d'un coup d'un seul les notions de site Internet BNUE, d'architecture technique sous-jacente, d'indexation des contenus par les moteurs grand public, de modalité de tri des résultats de recherche (« *page ranking* »), etc.
- Aux Bibliothèques nationales le choix de leurs contenus à numériser et à indexer par les acteurs retenus par l'entité commune.

Si l'on pousse le raisonnement à l'extrême, on pourrait penser que les gouvernements et Bibliothèques nationales européennes fondent une entité qui serait « OCA Europe ». Le concept séduisant et assez universel véhiculé par OCA n'appartient pas à ses fondateurs, et l'Internet s'est construit – et se construit toujours – ainsi , par appropriation de ses concepts universels et fondateurs par un groupe d'individus ou un groupe de nations : le nommage Internet (entité dans chaque pays), la gestion des serveurs racine, les encyclopédies libres en ligne. A titre de comparaison, l'Europe s'est saisie de la gestion des noms de domaine en .eu et l'a obtenue de la structure privée américaine en charge des serveurs racine de l'Internet (l'ICANN).
L'avantage de positionner dès le départ cette entité commune comme étant OCA Europe est :

- ✓ Se mouler dans le modèle Internet, où si un concept est jugé bon (c'est le cas d'OCA), il est « récupéré » par d'autres acteurs.
- ✓ Démarrer à partir d'un modèle existant, à savoir ce que fait OCA au niveau américain.
- ✓ Avoir une communication claire autour d'un concept « OCA Europe », plutôt qu'autour d'une BNUE aux contours encor flous. Cette communication qui asseoit l'action a une valeur externe, pour la presse et le grand public, mais a aussi une valeur interne, pour « resserrer les rangs » des Bibliothèques nationales européennes autour du projet.
- ✓ Etre en mesure de discuter simultanément avec Yahoo ! (membre fondateur d'OCA) et MSN (qui a rejoint OCA par la suite), avec peut-être la possibilité d'inclure l'accord MSN-British Library dans OCA Europe.

Comme nous l'avons fait pour la Commission Européenne (cf. ci-dessus), il convient de tempérer ce qui précède par le fait que cette stratégie repose quasi-exclusivement sur les Bibliothèques nationales des pays européens ; or, force est de constater qu'à ce jour, nous n'avons pas de vision claire de leur positionnement sur ce type de projets, et qu'à cet égard des contacts suivis doivent être pris avec elles suite à la lettre du 6 décembre 2005 du président de la BnF à ses homologues européens.

[17] Le secteur privé de l'édition pourrait aussi avoir cette même démarche commune au niveau européen.

Fiche 13 : Un nécessaire recentrage des politiques de soutien, en France comme en Europe

Si le projet BNUE est amené à prendre une certaine ampleur, il est nécessaire de voir dès maintenant comment une certaine coordination doit être menée au niveau des divers plans de numérisation et des diverses réalisations de bibliothèques numériques, en France comme en Europe.
Il ne s'agit pas de mettre un terme à des projets qui vivent leur vie, souvent s'adressant à des publics différents (p.e. les divers portails français soutenus par l'Etat en revues de Sciences humaines et sociales), mais de savoir ce qui existe, afin qu'il n'y ait pas double numérisation ou double développement informatique, tout en s'inspirant pour la BNUE de certains outils (notamment affichage, moteurs,…).

France- réalisations en cours.

Nous pouvons décrire l'activité française en la matière suivant deux axes :

1. La numérisation et l'accès aux Revues de sciences humaines et sociales, avec des niveaux d'érudition des contenus qui ne sont pas forcément prioritaires pour la cible de public BNUE.
   On peut recenser en ce domaine pas moins de quatre bibliothèques numériques soutenues financièrement par les pouvoirs publics :
   - Le portail www.revues.org, mis en œuvre par l'EHESS (Ecole des hautes études en sciences sociales) avec le soutien du CNRS et du Ministère de la Recherche.
   - Le portail Adonis du CNRS, http://edition.cens.cnrs.fr/revue/
     Annoncé en février 2005, ce portail a été ouvert en décembre 2005 avec un premier échantillon de revues partenaires (Bulletin du centre de recherche français de Jérusalem, Cahiers Charles V, Nouveaux cahiers d'allemand,…)
     Initiative du CNRS, il a reçu le label de « **Très Grand Equipement** », label gouvernemental accordé à des projets désignés comme prioritaires (Ministère de la Recherche)
     Structure : UPS créée le 15 janvier 2005 auprès de l'ENS Lyon.
     Budget et personnel : 34 M€ sur 10 ans, une centaine de chercheurs et ingénieurs
     Caractère européen : une version internationale du TGE ADONIS est à l'étude dans le cadre des institutions européennes et des contacts bilatéraux sont également en cours avec les grands organismes de recherche des pays à forte tradition dans les Sciences de l'Homme et de la Société : Grande Bretagne, Allemagne, Italie, Espagne, Hollande, Canada, États Unis (*source http://info.cens.cnrs.fr/article8.html*)
   - Le portail www.cairn.info, société de droit privé belge, avec des éditeurs français (Belin, La Découverte,…), et soutenue par la participation de la BnF au capital (voir fiche sur le partenariat public-privé)
   - Le portail www.persee.org lancé par la Direction de l'Enseignement Supérieur en 2004 (cf. démonstration au Comité de pilotage du 30 août).

   Il est certain que ces différents portails de revue SHS pourraient bénéficier d'une meilleure coordination ; toutefois ils s'adressent en priorité à des chercheurs qui

font appel de manière ciblée à ces sites qu'ils connaissent ; en ce sens, ces sites s'adressent à des publics plus restreints que ceux visés par la BNUE. Parmi ces sites, c'est le portail Persée qui paraît le plus proche du point de vue technique et des contenus de ce que nous souhaitons voir dans la BNUE.

2. Les plans de numérisation de la puissance publique, via l'Etat (établissements publics comme la BnF, ou Plan national de numérisation du Ministère de la Culture, ou programmes de soutien MENESR dont ceux figurant ci-dessus) ou via les Collectivités locales (bibliothèques numériques des grandes bibliothèques municipales).
   Le groupe GT5 (Financements) du Comité de pilotage BNUE s'est penché sur ces différents plans, et a établi le tableau suivant en matière de financements publics :

Organisme	Programme	Bénéficiaires	Objectifs	Montant annuel (€)	Résultats/Commentaires
Ministère de la Culture	Plan de numérisation	Bibliothèques, musées, archives Etat	Diffusion du patrimoine culturel : financement et coordination des programmes de numérisation ; formation des personnels ; catalogue national des projets de numérisation (numerique.culture.gouv.fr)	1,2 M	Bases nationales d'images (Musées, Archives) : diffusion systématique des documents ; mise en ligne du catalogue, intégration dans projet européen Michael (puis Michael +), à terme catalogue de liens vers les documents numérisés
		Bibliothèques (15%), musées, archives Collectivités territoriales (partenariat avec les collectivités)		1,5 M	Mises en ligne posent parfois problème dans les bibliothèques notamment. Fonds anciens numérisés. Impulsion actuelle du MCC en direction de l'écrit (Presse)
	Subventions aux grands établissements	INA	Numérisation des collections audiovisuelles	(10 M)	Exclu du total car audiovisuel uniquement
		BNF	Gallica Dont incitation à numérisation partagée avec Pôles associés	1,8 M (hors emplois)	80 000 volumes de texte, 80 000 images en ligne
	Subvention à des équipes de recherche (partenariat avec CNRS)	Institut de recherche et d'histoire des textes	Numérisation de manuscrits	71 000	
		Centre d'études supérieures de la Renaissance	Incunables	A instruire	
		Direction des archives de France		A instruire	
Centre national du Livre	Aide aux revues en ligne	Editeurs	Développement des revues électroniques (ex : CAIRN)	1 M	Source du financement : photocopies
Ministère de l'éducation nationale	Actions spécifiques	Consortium	Persée (portail de revues en SHS)	580 000	Apport des universités sur leurs ressources propres, budget total = 1,8M
	Actions spécifiques	BIUM+CNAM	Portails thématiques	520 000	
	Contrats avec établissements	Bibliothèques universitaires		120 000	Critères d'éligibilité : corpus, accessibilité ; tend vers l'harmonisation des formats (OAI-PMH) ; travail en partenariat (portails régionaux type NordNum)
Ministère de la Recherche		Cellule MathDoc	Numérisation et indexation d'articles	63 200 (fonctionnement) + 23 000 (1 salaire) + 2 salaires	Financement CNRS+Université
		Autres laboratoires CNRS		A instruire	
Collectivités territoriales (Plan Etat-Région type Banque numérique du savoir d'Aquitaine, Programmes régionaux type Rhône-Alpes, Municipalités)		Bibliothèques municipales, centres d'archives, musées		A instruire	
Total identifié				**6 877 200**	

Figure 13.1 Tableau des financements publics consacrés à la numérisation en France en 2004 (élaboré par le GT5 Financements du COPIL)

Nous voyons à travers ce tableau une dispersion certaine d'un montant de 6,9M€ somme toute limitée.
Une coordination entre d'une part le projet BNUE, avec les montants qui lui sont affectés de 0,4M€ en 2006, et d'autre part le plan national de numérisation du MCC, doit être envisagée, au moins pour la partie « imprimé » du plan MCC.

Commission européenne- The European Library (TEL) et Michaël

Au niveau de la Commission Européenne, c'est une grande dispersion des fonds et des projets sur la numérisation et les bibliothèques numériques qui frappe.
La note communiquée par le Ministère de la Culture (Direction au Développement et aux Affaires Internationales) au Comité de pilotage du 30 août 2005 montre le foisonnement de projets et de sources de financement existant actuellement en Europe.
Les programmes de recherche (type IST dans le PCRD) sont déjà une source de financement importante pour des sites de bibliothèques numériques expérimentaux ou à public très ciblé : on relèvera par exemple le projet DELOS http://www.delos.info « Network of Excellence on Digital Libraries ».

Nous avons choisi de nous intéresser à deux projets européens spécifiques, en ce sens qu'ils recouvrent très largement certains objectifs de la BNUE. Nous n'avons à ce stade malheureusement pas pu faire figurer les montants de financements européens impliqués dans ces deux projets.

1. Le projet Michaël, Multicutural Inventory of Cultural Heritage in Europe, cf. http://www.michael-culture.org/index_f.html

Extrait de la présentation de Michaël

Le projet Michael (inventaire multilingue du patrimoine culturel européen) est un travail commun élaboré entre la France, l'Italie et le Royaume-Uni, soutenu par la Commission européenne - programme eTen, dédié au déploiement des nouvelles technologies en Europe.
Michael vise à proposer un accès simple et rapide aux collections numérisées des services du patrimoine, des musées, des bibliothèques et des archives des différents pays européens. Cette plate-forme multilingue est dotée d'un moteur de recherche capable de restituer les informations sur des collections dispersées dans des lieux et sur des serveurs différents. Les applications sont nombreuses, pour l'éducation et la recherche, mais aussi pour le développement de services commerciaux innovants, notamment pour le tourisme et l'éducation. Cet outil constitue donc une étape indispensable vers la création d'un espace culturel européen sur Internet.
Par ailleurs, Michael permet de valoriser non seulement les collections nationales, mais aussi les savoir-faire en matière de numérisation et de constitution d'inventaires numérisés. De cette façon, il occupe une place de premier plan dans la politique de développement de services en ligne en Europe.

Ce projet se différencie de la BNUE, au sens où il va au-delà du texte imprimé, (archives, images, audiovisuel,…), et surtout où il propose un accès aux catalogues (comme TEL ci-dessous), et non aux œuvres en direct : ce simple objectif paraît très limité, eu égard aux fonds par ailleurs investis par la Commission Européenne (3M€ à ce jour), et à l'absence de notoriété du site pour le grand public.

2. Le projet TEL, The European Library, http://www.theeuropeanlibrary.org

> ***Un précurseur européen défunt : la Bibliotheca Universalis***
>
> *Au cimetière des projets transnationaux, nous devons rappeler la Bibliotheca Universalis, présenté comme « projet pilote du G7 sur la société de l'information en 1994 »* (http://www.ifla.org/IV/ifla64/031-98f.htm)*, associant au départ la France et le Japon.*
>
> *Ce projet, malgré les fonds investis, n'a jamais vu le jour ; la seule adresse Internet connue qu'il ait*[18] *pointe maintenant vers le projet communautaire qui a pris son relais, The European Library.*
>
> *Comme le souligne la note de la Direction du Livre transmise au Comité de pilotage du 30 août : « En 1999, treize bibliothèques participaient à ce projet aujourd'hui hébergé par la Bibliothèque royale des Pays-Bas, sur Gabriel, le serveur de la Conférence européenne des bibliothèques nationales (CENL)*[19]*. Le projet Bibliotheca universalis n'a abouti à ce jour qu'à des résultats modestes (site web, essai d'orientations communes, colloques…) ».*

Actuellement, et ce depuis novembre, date à laquelle le connecteur OAI a pu être mis en place avec la BnF, ce site répertorie principalement des contenus numériques Gallica (10 000 ouvrages sur les 80 000 de Gallica).
Une recherche sur un auteur donne :

- La liste des notices bibliographiques dans l'ensemble des bibliothèques nationales européennes (sans accès aux œuvres).
- La liste des documents numérisés, qui se résume à deux fonds européens :
  - ✓ Le fonds Gallica.
  - ✓ Le fonds italien Bibliothèque de Florence, avec deux inconvénients majeurs sur ce fonds : seule la couverture des livres est visible (!), et dans un visualisateur absolument non convivial (programme à télécharger). Certains livres sont visibles de manière payante.
  - ✓ Ce dernier exemple montre que le chemin est long pour une interface BNUE conviviale commune.

Dans TEL, il n'y a pas visualisation commune, on est exactement dans le cas de figure d'architecture 1.2/3.3 (voir fiche architectures), totalement décentralisée,

[18] http://www.kb.nl/gabriel/bibliotheca-universalis/en/bibliotheca_universalis_accueil.htm (cette adresse est répertoriée par exemple sur la page du site de la Bibliothèque Nationale d'Espagne consacrée au projet, http://www.bne.es/esp/labicobibliotheca.htm).
[19] www.kb.nl/gabriel/bibliotheca-universalis/fr/bibliotheca_universalis_projet.htm

où l'internaute est envoyé dans l'environnement Gallica pour voir un document Gallica.

Le projet est managé par la bibliothèque Royale des Pays-Bas, dans un bureau qualifié de « Bibliothèque européenne » (qui correspond à la traduction de « European Library » (voir http://libraries.theeuropeanlibrary.org/contactus_fr.htm)

On peut s'étonner du faible niveau d'avancement de ce projet (et ce d'autant qu'il faisait suite au projet Bibliotheca Universalis lancé en 1995, et malgré les financements investis), et du caractère totalement confidentiel et absolument pas orienté vers les internautes de cette réalisation.

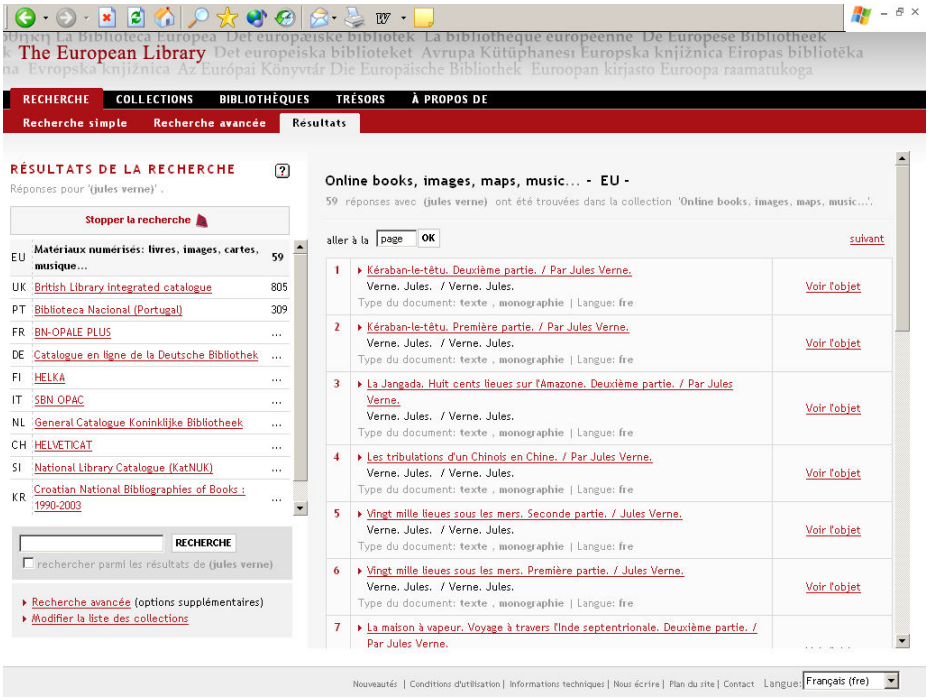


Figure 13.2 : Par exemple sur une recherche Goethe sur TEL, on trouve 155 ouvrages dont 120 sur Gallica (dont certains en allemand…) et le reste sur le site Bibliothèque de Florence.

Fiche 14 : Deux sujets complémentaires importants pour le projet

Ces deux sujets ne sont pas directement liés au plan d'actions immédiat de la BNUE, mais sont fondamentaux à long terme dans l'esprit du déploiement d'une telle bibliothèque.

Préservation des contenus numériques.

Ce sujet, même s'il n'est pas le plus crucial sur le moment, est important à long terme. Chacun a en tête des exemples de programmes informatiques dont on a besoin des codes-sources et qu'on n'arrive plus à lire sur les ordinateurs actuels.
Bien évidemment l'on peut faire des sauvegardes régulières, mais le sujet reste posé de la longévité des supports physiques comme le CD-Rom.
Les grands acteurs de l'Internet ne s'y sont pas trompés, qui ont conclu des accords de recherche avec de grandes institutions : par exemple Microsoft, en même temps que l'accord de MSN avec la British Library, a conclu un accord de recherche sur la préservation des contenus numériques entre son laboratoire de recherche européen basé à Cambridge (UK) et l'University of Cambridge.

Aux Etats-Unis, la Library of Congress bénéficie depuis 2000 d'une dotation pluriannuelle de 99,8M$ sur la numérisation : il s'agit d'un programme national appelé le National Digital Information Infrastructure and Preservation Program (NDIIPP), avec son site internet www.digitalpreservation.gov .
Dans ce cadre, après un appel à propositions en 2004, elle a annoncé en mai 2005, en collaboration avec l'Agence de recherche américaine la National Science Foundation (NSF), une dotation de 3M$ pour lancer des programmes dans dix universités américaines sur le sujet de la recherche en préservation des contenus[20].

Dépôt légal numérique.

Le fait de rendre à terme le dépôt légal numérique obligatoire en complément du dépôt légal papier aurait le double intérêt suivant :

- Jouer le rôle de préservation numérique de ces documents, rôle auquel pourraient s'engager les organismes de dépôt légal en Europe. Un fichier numérique stocké à la BnF pourrait être retrouvé 70 ans plus tard, pas forcément chez son éditeur (cf. problème des œuvres orphelines, par exemple)
- Préparer la Bibliothèque Numérique de nos petits-enfants, qui comprendraient difficilement ne pas trouver trace de fichiers numériques à partir de 1995, date à laquelle la quasi-totalité des livres imprimés l'a été à partir d'un format électronique natif !

[20] Voir liste des dix sujets de recherche à http://www.digitalpreservation.gov/about/pr_050405.html


FIN