

PRESERVER LES CONTENUS DU WEB

Julien MASANES

Bibliothèque nationale de France, Paris, France

4ème journées internationales d'études de l'arsag

RESUME

Préserver les contenus du Web nécessite de redéfinir les méthodes de travail pour le repérage, la collecte et le stockage de ce type d'information. Cela implique également de réaliser un certain nombre d'actions techniques en vue de pérenniser les contenus pour garantir la possibilité d'un accès à long terme.

Une présentation des spécificités de cet espace documentaire est proposée et les implications pour la conservation sont abordées. La difficulté particulière que représente la préservation des sites dynamiques est notamment envisagée en détail avec la présentation d'un modèle de migration des sites utilisant des bases de données relationnelles comme passerelle documentaire, type de sites relativement fréquents et souvent riches sur le plan documentaire.

Des exemples tirés des expérimentations en cours à la BnF sont présentés.

INTRODUCTION

La nécessité d'organiser une préservation des contenus du Web, vecteur incontournable de diffusion du savoir et de la culture de notre époque, apparaît aujourd'hui de manière évidente.

La conservation de ces contenus pose une série de problèmes dont nous n'aborderons ici qu'une partie. Parmi ceux que nous n'aborderons pas, citons pour mémoire le problème du repérage, de la collecte et du stockage des grands volumes d'information présents sur le Web.

Nous nous focaliserons ici sur le problème de la pérennisation des archives de sites Web, pour les sites statiques comme pour les sites dynamiques qui forment ce que l'on appelle fréquemment le 'deep Web' et qui tendent à se généraliser aujourd'hui.

Préserver un site c'est bien sûr préserver un ensemble de fichiers mais c'est aussi préserver la structure de cet ensemble et son mode de fonctionnement, notamment la navigation qui est le mode naturel d'accès à ce type de contenus. Cela s'avère particulièrement ardu lorsque ce fonctionnement est dépendant d'une architecture technique complexe comme c'est le cas pour les sites dynamiques. Pour illustrer ce problème et les manières de tenter de le résoudre, nous prendrons des exemples tirés des expérimentations en cours à la BnF et notamment la réalisation d'un démonstrateur de migration des sites dynamiques.

LE WEB COMME ESPACE DOCUMENTAIRE

Avant d'aborder la question de la conservation des contenus du web, il convient de caractériser cet espace documentaire du point de vue de la localisation et de la structuration particulière de contenus qu'il introduit.

LES SITES STATIQUES

Internet est un réseau de réseaux donnant accès eux-mêmes à des serveurs. Sur ces serveurs sont stockés des fichiers avec une organisation locale en système de fichiers. Ces systèmes de fichiers sont organisés en arborescence de répertoires, propriété importante pour notre propos.

Un des protocoles les plus populaires de l'Internet est celui qui sert à mettre en ligne des publications, essentiellement sous une forme hypertextuelle mais pas uniquement. Sur le Web, l'accès aux serveurs de fichiers se fait au travers d'un serveur particulier gérant le protocole spécifique (HTTP) d'accès aux fichiers.

Fédérant l'ensemble de ces systèmes de fichiers locaux, le web peut être considéré comme un vaste système de fichiers global auquel on accède par la navigation hypertexte. Les liens pointent sur des fichiers qui une fois chargés sont affichés comme des pages. Ces pages, présentes sur les serveurs sont appelées des pages statiques, par opposition aux pages dynamiques qui ne sont construites qu'à la demande, lors de la navigation.

Sur les sites statiques, la construction des adresses se fait par l'ajout d'un niveau supplémentaire à l'arborescence correspondant à l'adresse locale.

Ex :

Adresse du fichier.html sur le serveur local :

Disque 2/www/repertoire/sous-repertoire/fichier.htm

qui devient sur Internet cette adresse

http://www.site.fr/repertoire/sous-repertoire/fichier.htm

Ce qui nous intéresse ici est de noter que l'arborescence du site est, sans prendre en compte la première partie de l'adresse, la même que l'arborescence du serveur.

Un autre niveau d'appréhension de la structure d'un site, le niveau logique de la navigation hypertexte est distinct de ce niveau de stockage physique des fichiers, tout en lui étant étroitement articulé par le mécanisme des liens hypertexte propre à HTTP (voir figure 1). En cliquant d'une page à une autre, l'internaute se déplace en réalité dans le système de fichier du serveur, avant de le quitter pour celui d'un autre serveur lorsqu'il quitte ce site pour un autre. Toute l'information d'aiguillage dans les nombreux fichiers qui forment un site est contenue dans les liens hypertextes qui renferment l'adresse réelle du fichier. Cela permet à l'internaute de ne pas se soucier de cette adresse et de passer d'un répertoire à l'autre, d'une machine à l'autre, voire d'un continent à l'autre sans s'en rendre compte.

Les liens hypertextes d'un site forme donc une grille permettant l'articulation étroite du niveau logique de navigation avec celui, physique, de stockage des fichiers.

Double pointage logique et physique des liens hypertextes

- - - ➔ Déplacement d'un document à l'autre par un clic sur les liens

— ➔ Demande de document envoyée au serveur avec l'adresse réelle du fichier

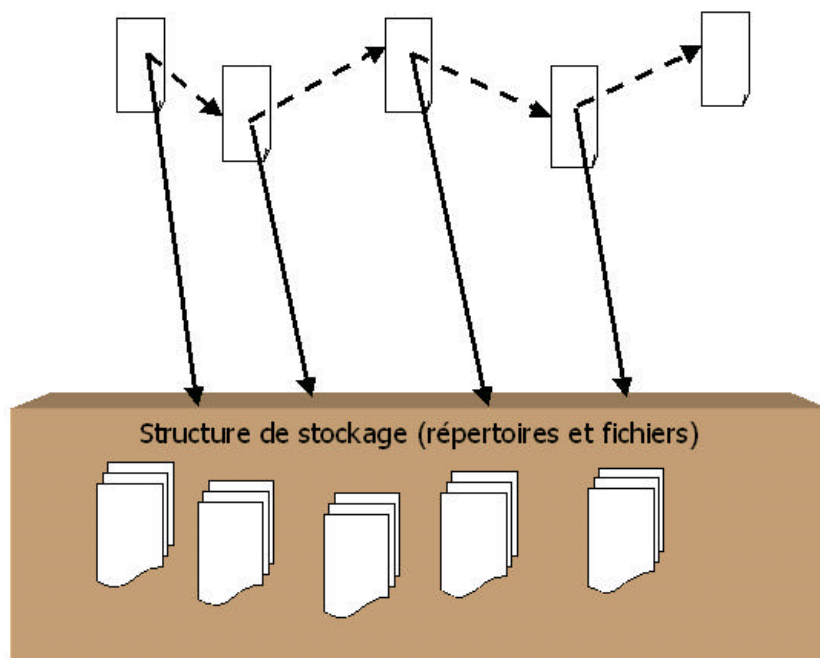


Figure 1

LES SITES DYNAMIQUES

Le modèle de sites précédent n'est plus aujourd'hui le seul ni même le principal. Nombre de sites que l'on appelle dynamiques utilisent des bases de données pour stocker une partie ou la totalité de leur contenu. La différence essentielle tient à ce que les pages telles qu'elles sont visualisées côté client, n'existent pas sur le serveur. Elles sont générées à la volée lorsque la requête est envoyée aux serveurs (voir figure 2). Le serveur HTTP transmet aux serveurs d'application (qui permettent d'exécuter les programmes) les paramètres de la requête et les scripts ou programmes génèrent une page HTML à partir de données dont une partie peut être extraite de bases de données.

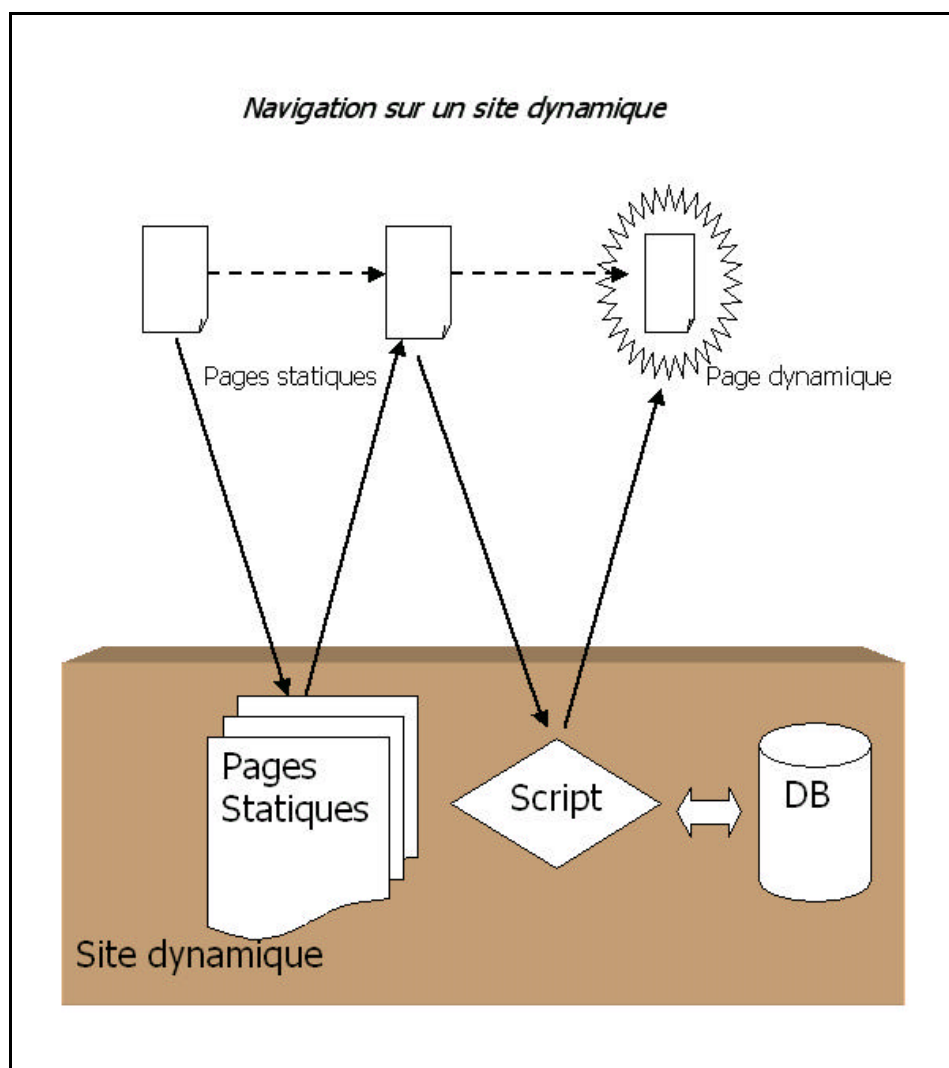


Figure 2

On observe deux grandes catégories de sites dynamiques utilisant de telles architectures. La première consiste à stocker dans une base de données le contenu de page ayant toutes la même structure (par exemple des fiches descriptives ou des messages de forum) . Cela évite d'écrire pour chacune les éléments de structure des pages. Ces éléments sont générés par un script qui se charge de remplir la page envoyée au client avec le contenu stocké dans la base.

De telles architectures (que j'appellerai du « HTML éclaté » dans des Bases de Données Relationnelles ou BDR) sont mises en place essentiellement pour faciliter la gestion du site, notamment les mises à jour.

L'autre grande catégorie d'utilisation d'architectures dynamiques se rencontre lorsque le site sert de passerelle d'accès à des espaces documentaires non-Web. Les bases de données servent alors soit à stocker directement l'information visée, comme dans le cas de données scientifiques mises en ligne, soit à obtenir les informations qui vont permettre d'accéder aux documents visés par exemple un catalogue d'articles ou de photos. Ce dernier cas, que j'appellerai celui des « passerelles documentaires », se retrouve fréquemment dans les grands sites formant des réservoirs documentaires sur le Web.

Ces deux types de sites dynamiques (« HTML éclaté » et « passerelles documentaires ») utilisent des modes d'accès aux contenus différents. Le premier type utilise la navigation hypertexte. Les liens ont ceci de particulier qu'ils servent à passer les paramètres utilisés pour la génération des pages. Ils sont formés d'une partie classique contenant l'adresse du programme ou script et d'une seconde partie (après un point d'interrogation) qui contient des paramètres sous la forme 'attribut = valeur'.

<http://www.site.fr/repertoire/monscript.pl?nom=Hugo&prenom=Victor>

Dans cet exemple, cliquer sur ce lien déclenchera le script appelé 'monscript' et écrit en Perl (d'où le '.pl') avec les paramètres nom=Hugo et prenom=Victor. Ce script pourra par exemple créer et afficher une fiche biographique dudit Victor en allant chercher l'information dans une base de données biographique.

Le second type, les passerelles documentaires, peut également utiliser ces liens dynamiques. Mais le plus souvent, il est nécessaire que l'utilisateur spécifie les informations concernant le document visé (par exemple l'auteur d'un article ou le thème d'une photo). C'est alors la technique du formulaire qui est utilisée, permettant à l'internaute de spécifier les caractéristiques de l'information ou du document dans des champs, dont le contenu est envoyé après validation au serveur.

(voir figure 3)

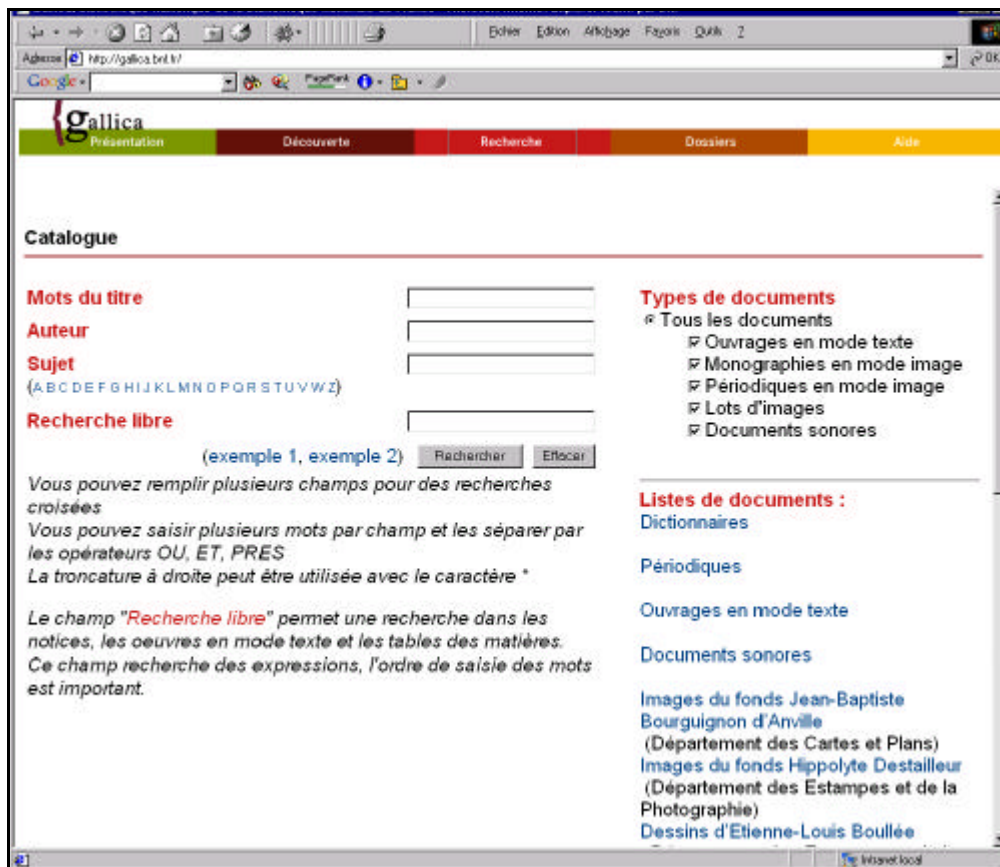


Figure 3

CONSERVATION A LONG TERME

LES SITES STATIQUES

La préservation des sites statiques ne pose pas de problèmes insurmontables. Il est possible « d'aspirer » ces sites en ligne ou de faire des copies sur un serveur d'archive. Dans les deux cas, le plus simple est de reproduire dans l'archive l'arborescence telle qu'elle existe sur le serveur. Cela permet d'accéder au contenu par une simple navigation.

Si les pages HTML du site contiennent des liens absolus (contenant l'ensemble de l'adresse), la navigation ne sera cependant pas possible, puisqu'un clic sur ces liens enverra à l'adresse du site réel. C'est pourquoi il est préférable de transformer les liens absolus en liens relatifs (qui ne contiennent qu'une partie de l'adresse, et prennent pour point de départ le point où l'on se trouve) lors de l'archivage. Cela permet d'assurer simplement la possibilité de naviguer sur l'archive à quelque endroit que celle-ci soit stockée, puisque les liens 'fonctionnent' alors par rapport à l'adresse de stockage.

Pour un stockage à long terme, il est également nécessaire de s'assurer de la pérennité des formats utilisés. De fait, on constate une forte présence sur le Web de formats dont les spécifications sont publiques (HTML, XML, JPEG, GIF, PNG, Flash, PDF, RTF etc.) ce qui permet d'être assuré de la possibilité de les lire à l'avenir.

LES SITES DYNAMIQUES

La préservation de sites dynamiques pose beaucoup plus de problèmes car leur contenu n'est pas récupérable directement puisque construit lors de la consultation. Ces sites dépendent d'une architecture complexe, faisant intervenir différents systèmes dont des bases de données.

La conservation à long terme de ces différents systèmes, des paramètres et des connexions qui leur permettent de travailler ensemble est utopique.

La conservation des bases de données à elle seule est déjà délicate. En effet les bases de données sont dépendantes d'un moteur propre, lui-même dépendant d'un environnement informatique donné et historiquement daté.

Il est donc nécessaire d'effectuer une sorte de mise à plat des sites dynamiques pour pouvoir les conserver.

Pour effectuer une telle mise à plat, la méthode est relativement simple pour les sites dont les pages sont éclatées dans des bases de données (« HTML éclaté »). Il est en effet possible dans ce cas de générer toutes les pages du site en utilisant un robot automatique qui va parcourir tous les liens qu'il va trouver sur le site et stocker les pages ainsi générées. L'archive sera donc uniquement composée de page HTML et le problème de la conservation des bases de données et de leur environnement ne se pose donc plus. Tout le contenu est présent dans des fichiers statiques. Il reste à l'archiver et nous sommes ramenés au cas précédent (sites statiques).

La préservation des passerelles documentaires

Pour les sites utilisant des bases de données comme passerelles documentaires la question est plus complexe.

On peut découper schématiquement ces sites en trois parties (voir figure 4).

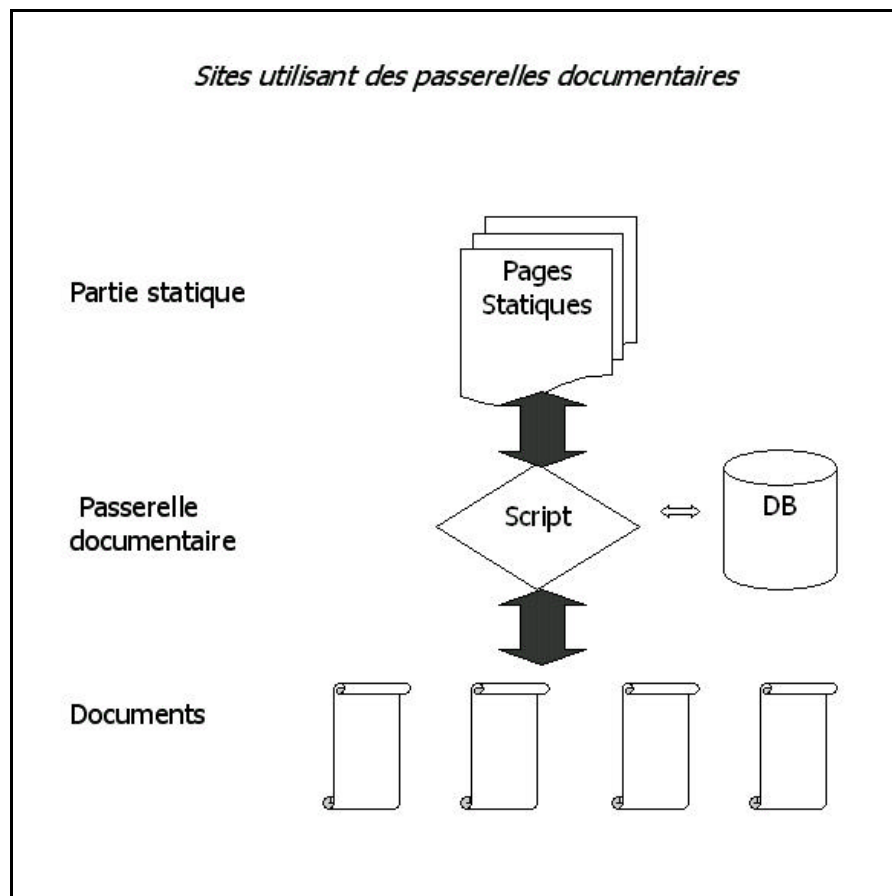


Figure 4

La première est la partie éditoriale souvent en HTML, qui se ramène au cas des sites statiques ou à la première catégorie de sites dynamiques.

La deuxième partie est constituée des scripts et des bases de données servant de passerelles documentaires.

La troisième partie, ce sont les documents eux-mêmes, auxquels on accède à travers cette passerelle.

La conservation de cette dernière partie dépend essentiellement du format des documents, et de sa pérennité. Il peut être opportun de réaliser une migration lors de l'archivage si besoin est.

La réelle difficulté vient de la deuxième partie, celle des passerelles documentaires. L'information contenue dans ces bases joue en effet un rôle essentiel pour l'accès aux documents. Si l'on veut assurer la conservation d'un site qui contient des dizaines de milliers d'images, il est absolument nécessaire de conserver les moyens d'accéder facilement à cette masse de documents. L'identifiant du document, ses caractéristiques, sont contenues dans les bases de données formant la deuxième partie du site.

On ne peut se contenter d'archiver le formulaire de recherche et les documents. Il est nécessaire de conserver la passerelle documentaire en état de marche.

La stratégie de préservation testée dans le cadre des expérimentations actuelles à la BnF a pour objectif la migrer ces passerelles vers un format, le XML¹, qui tout en étant pérenne, permet des fonctionnalités de recherche d'information. L'utilisation d'une archive en format semi-structuré (XML) permet en effet de garantir un accès futur simple à cette information. Contrairement aux bases relationnelles dépendantes d'un logiciel spécifique, XML est une syntaxe entièrement documentée et accessible indépendamment d'un environnement informatique donné (on peut à la limite le lire directement sur une sortie papier).

Cette syntaxe offre un niveau d'organisation de l'information équivalent aux BDR et permet donc d'effectuer des recherches sur des portions d'information (correspondant aux champs des BDR). Certes les performances de la recherche en terme de rapidité ne peuvent être comparées aux Systèmes de Gestion de Base de Données Relationnelles (SGBDR). Cela est cependant largement compensé de notre point de vue d'institution de mémoire par la possibilité de pérenniser une partie essentielle de l'information contenue sur ces sites.

Un démonstrateur est en cours de test à la BnF sur deux sites utilisant des SGBDR comme passerelle documentaire. Un site de vente de livres électroniques et un site contenant un nombre important d'images médicales.

L'accès à l'archive du site se fera par une navigation traditionnelle (voir figure 5) sur la partie statique du site et la partie passerelle sera remplacée par une mécanique d'aiguillage propre mais navigable. Elle comprendra notamment un formulaire, différent du formulaire d'origine, permettant d'interroger la base XML et aiguillant vers les documents stockés.

¹ XML est une syntaxe normalisée permettant de définir une structuration d'informations à base de balises. C'est un descendant de SGML et il est normalisé par le W3 consortium, organisme développant les standards du Web.

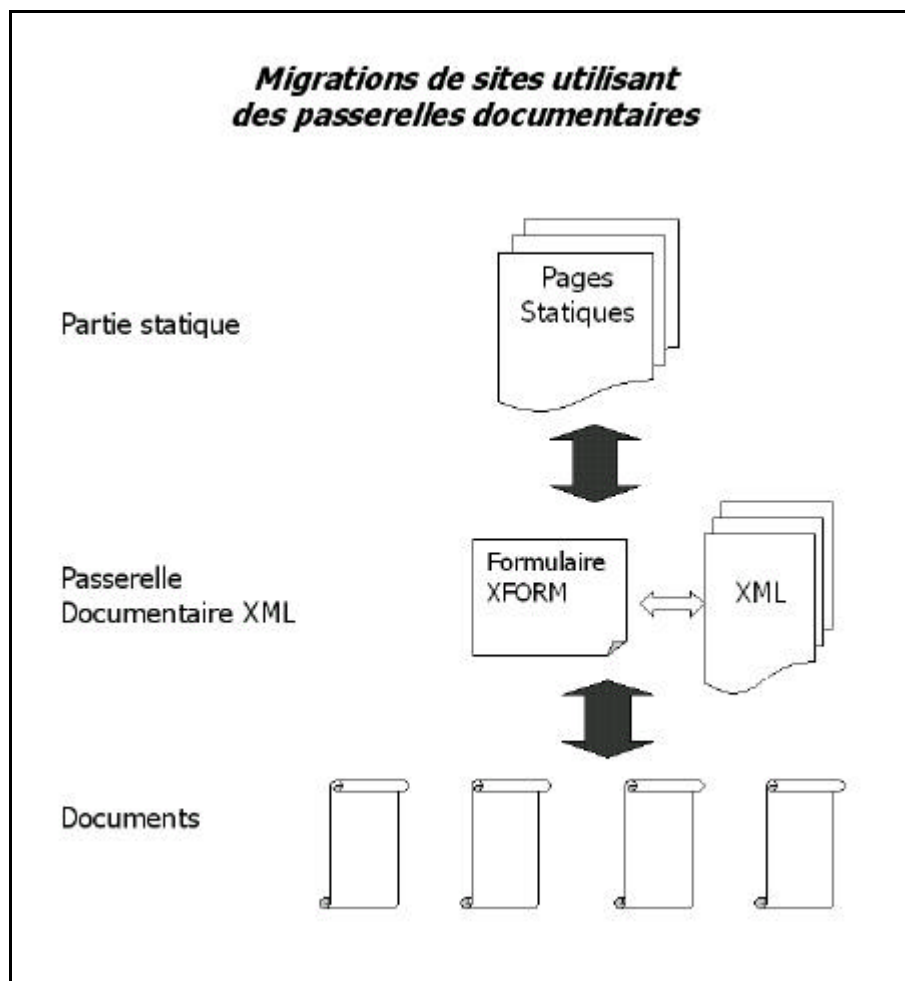


Figure 5

L'accès à l'archive ne sera ainsi pas conservé sous sa forme originelle exacte mais la navigation sera toujours possible et cela sur une couche archivée pérenne.

Voyons plus en détails ce modèle.

Modèle de migration des passerelles documentaires 'relationnel vers semi-structuré'.

La première difficulté à laquelle on se trouve confronté est celle de l'hétérogénéité des bases documentaires présentes dans les sites. Les champs de ces bases, même lorsqu'ils portent sur une même entité comme l'auteur, portent rarement le même intitulé. Leur type peut également différer. Vu la masse de sites à traiter (plusieurs centaines), il est exclu d'adapter les outils à chaque cas particulier. Il est nécessaire d'aboutir à un niveau d'information documentaire générique permettant d'appliquer les mêmes outils (formulaire, index, mécanique de pointage vers les documents) dans tous les cas.

Heureusement, il s'avère sur plusieurs dizaines de sites examinés que les catégories d'information utilisées pour la recherche documentaire sont limitées et que définir ce niveau générique n'est pas hors de portée.

Le niveau de description simple que propose le Dublin Core n'est en fait jamais entièrement utilisé. Pour les cas que nous traitons, le niveau de description minimal ('Auteur', 'Titre', 'Mots-clés', éventuellement 'Période') s'avère largement suffisant.

La définition d'une structure commune d'information est ainsi possible autour de ces catégories générales d'information.

On constate également qu'une partie seulement de l'information contenue dans ces bases est utilisée pour l'aiguillage documentaire sur le site.

On peut donc éliminer toute la partie de l'information de ces bases qui ne fait pas à proprement parler partie de la publication (information interne de gestion des documents notamment).

Pour le test en cours à la BnF, trois catégories d'information ont été retenues : 'Auteur', 'Titre', 'Mots-clés'.

Le schéma XML générique est construit à partir de ces trois champs.

La migration passe par une mise en correspondance des champs de la base d'origine avec ces entités. Cette phase, nécessitant une bonne connaissance de la structure de la base et de son contenu pourrait être effectuée en ligne par les producteurs de la base lors du dépôt. Dans cette phase de l'expérimentation, elle sera réalisée à la BnF à partir des informations données par les producteurs.

Pour chaque enregistrement de la base, un document XML est généré respectant le schéma générique. Les questions d'implémentation à grande échelle, notamment celle de savoir s'il serait préférable de ne générer qu'un seul document XML pour l'ensemble de la base, ne seront arbitrées que dans une phase ultérieure de l'étude.

Les documents XML générés seront stockés et indexés, certainement dans un SGBDR vu l'état des systèmes existant aujourd'hui. Mais l'important est qu'indépendamment de toute implémentation (et c'est bien le but recherché pour la préservation à long terme), un accès garantissant certaines fonctionnalités de recherche soit possible.

Un formulaire générique utilisant le langage du W3C Xquery sera utilisé pour remplacer le formulaire de recherche d'origine du site. Via ce formulaire, une recherche documentaire sera possible selon des modalités voisines de ce qui existait sur le site mais sur une couche d'information pérenne (XML).

Un lien pointant vers l'adresse du document archivé (ici photo ou e-book) sera affiché comme réponse à la requête, ce qui permettra de poursuivre la navigation vers le document.

Cette 'greffe' d'un niveau de recherche générique s'intégrera ainsi fortement dans l'architecture du site archivé. Elle permettra de garantir la continuité de la navigation sur l'archive au travers de passerelles documentaires dont la conservation en l'état serait une tâche irréalisable.

Certes, il s'agira d'un niveau artificiel de recherche et il sera nécessaire de l'indiquer clairement à l'utilisateur de l'archive. Il est également envisageable de garder une version du formulaire de recherche d'origine accessible parallèlement à titre d'information.

L'essentiel (l'accès simple au site archivé) sera ainsi préservé. Entrer le nom de l'auteur d'un livre électronique dans un formulaire générique plutôt que dans le formulaire d'origine n'est finalement qu'un moindre mal. Les réservoirs documentaires non-Web d'un site (collection d'images ou de e-book) seront ainsi accessibles par une recherche documentaire même simple.

Avantages de ce modèle

L'avantage pour la préservation à long terme est évident : la dépendance vis-à-vis des paramètres, des scripts et des environnements propres à chaque site d'origine sera supprimée. L'archive constituée sera accessible au travers de technologies génériques et d'une couche d'information (XML) pérenne.

Ce modèle de migration est conçu pour pouvoir s'appliquer à l'échelle de plusieurs centaines de sites. Il devra permettre à la BnF d'effectuer un archivage selon ces modalités de plusieurs centaines de sites nécessitant une préservation de leurs passerelles documentaires.

En effet, les outils utilisés (schéma générique, logiciel de migration, formulaire de recherche, entrepôt de données XML) seront communs à l'ensemble des sites traités de cette manière. Seule la mise en correspondance de l'information propre à chaque site (champs de la BDR) avec le modèle générique nécessitera un travail particulier. On peut cependant envisager que cette mise en correspondance soit faite par le déposant lui-même lors du dépôt.

Difficultés et verrous à lever

1. Diversité des SGBDR utilisés par les éditeurs de sites.

Même si l'on retrouve fréquemment certains SGBDR (MySQL, SQL Server, Oracle), il n'en reste pas moins qu'il faudra être capable de traiter la diversité des systèmes existant.

C'est la raison pour laquelle plusieurs possibilités sont envisagées pour réaliser l'extraction/migration d'une partie de la base d'origine vers le schéma générique XML.

Les SGBDR les plus récents offrent des fonctionnalités d'exportation qui peuvent s'avérer satisfaisantes. Il pourra être dans certains cas possible de faire fonctionner un module propre ajouté au SGBDR. Enfin pour les autres cas, une extraction en format d'échange standard CSV sera utilisée, avec une migration à partir de ce format.

2. Respect de la structure de la base d'origine

Cette question se posera notamment pour les bases comprenant plusieurs tables avec des champs communs.

3. Pointage vers les fichiers.

La possibilité de réutiliser facilement les pointeurs vers les fichiers contenant les documents est primordiale pour garantir l'accès simple au document (sans recherche supplémentaire sur les noms de fichiers). Il reste à vérifier si les pratiques en la matière sont assez homogènes pour permettre de conserver son caractère générique au système de migration.

4. Volumétrie

L'information migrée en XML est nécessairement plus volumineuse que sous sa forme relationnelle. Cela dépend évidemment de la taille moyenne des champs. De plus, des index XML devront être générés. L'étude de deux cas concrets avec le démonstrateur devrait permettre de faire des évaluations chiffrées sur cette question.

CONCLUSION

La préservation des contenus du Web pose des questions nouvelles dont une partie a été abordée dans cet article. La réflexion et les expérimentations dans ce domaine portent sur des aspects qui intéressent les autres domaines de la préservation numérique notamment la définition de stratégie et de modèle de pérennisation pour l'information contenue dans des BDR.

L'évolution du Web vers une généralisation des architectures complexes oblige les institutions de mémoire à conserver, pour pouvoir garantir un accès à long terme, un niveau 'à plat' d'archivage, le plus indépendant possible d'une génération d'équipement informatique.