Production des sites : les enjeux actuels

Olivier Roumieux

Si le web est aujourd'hui ce qu'il est, un média incontournable pour tout éditeur, petit ou grand, c'est notamment grâce à une formidable démocratisation de ses outils de production. Beaucoup de simples citoyens ont été présents sur la Toile avant la plupart des grandes entreprises. Certes, dans les premières années, au milieu des années quatre-vingt-dix, il fallait faire preuve d'entregent pour dénicher quelques mégaoctets d'espace d'hébergement. Ce n'est plus le cas aujourd'hui. La plupart des outils dont nous allons parler possèdent des versions soit disponibles à des coûts modiques, soit entièrement gratuite[1]. De nombreux professionnels ont d'ailleurs appris chez eux, en réalisant leur site personnel. Ainsi est née la figure fondatrice du webmestre ou « *webmaster* », homme-orchestre capable, d'un même mouvement, d'écrire les pages de son site, de concevoir la charte graphique et de nouer deux ou trois partenariats générateurs de clics.

Aujourd'hui, la « gestion de projets » reprend ses droits et la production des sites dépasse largement la simple mise à disposition de fichiers sur un serveur. Nous assistons à un éclatement des compétences métier nécessaires. Une évolution due à la place stratégique qu'a (ont) acquise le(s) site(s) au sein de l'entreprise, mais également au déplacement du centre de gravité du projet depuis la salle des machines vers le poste de l'utilisateur, qu'il soit producteur ou consommateur d'information. Les enjeux sont désormais plus ambitieux : les pages doivent être produites par des rédacteurs à partir de simples formulaires, les contenus en ligne gérés avec la même attention que

1 Par exemple, l'outil de mesure de fréquentation Xiti dispose d'une tarification dépendant du volume de pages auditées et de deux formules (Pro ou Expert) à partir de 45 euros par mois ; une version limitée mais fonctionnelle est même entièrement gratuite. De nombreux outils de gestion de contenu, particulièrement ceux du monde libre, sont téléchargeables sans coût.

l'information interne, les performances d'accès et la facilité de navigation doivent être privilégiées et l'on ne peut plus se permettre de dissocier la mesure de la fréquentation de l'efficacité de la production.

◆ 1. Au commencement étaient le statique… et le dynamique

1.1 La valeur sûre du statique

Fonctionnellement et historiquement, le premier modèle de publication consiste à écrire ses pages HTML (HyperText Markup Language) à l'aide d'un éditeur plus ou moins sophistiqué, puis de les transférer directement sur un serveur web dédié ou, plus généralement, mutualisé. Cette organisation convient bien aux petits sites conçus par une ou deux personnes avec une centaine de fichiers. L'administration simple et légère permet de lancer un site très rapidement, sans travaux préparatoires excessifs. Le revers de la médaille, c'est que les sites ainsi conçus présentent fréquemment une arborescence sauvage, mal pensée et qui ne prépare guère les évolutions futures. Aucune collaboration entre contributeurs ne doit être envisagée. L'édition étant directe, il n'est pas possible de faire de tests en ligne avant publication définitive. Cela reste malgré tout le modèle adapté pour un prototypage rapide sans que l'on ait besoin d'investir avant le lancement d'un projet.

Peu d'outils suffisent pour démarrer : un éditeur HTML tel que Dreamweaver ou FrontPage, un logiciel de création graphique (Photoshop, Paint Shop Pro, etc.) et éventuellement un utilitaire de transfert par FTP (File Transfer Protocol) si l'on souhaite contrôler plus finement ses mises en ligne.

Quand le site prend de l'importance, en termes de fichiers et de contributeurs, il devient intéressant d'installer un serveur intermédiaire de pré-diffusion. Les modifications sont d'abord transférées sur ce serveur et il devient ainsi possible de tester la cohérence de différentes parties ou de la totalité du site avant publication. Ce serveur intermédiaire permet également de constituer une première chaîne de validation : le responsable éditorial valide toutes les modifications en consultant la version en pré-diffusion avant d'autoriser la diffusion sur le serveur ouvert aux internautes.

Dans la pratique, toute modification globale du site doit être répercutée manuellement sur chaque page du site. Pour pallier ce problème, outre la fonction bureautique « Rechercher / Remplacer », les éditeurs HTML les plus avancés tels que Dreamweaver proposent un début de factorisation de la production avec les modèles de pages et les bibliothèques d'objets. Si l'on

dispose d'un modèle de page par rubrique, il suffira de modifier ce seul modèle pour répercuter les changements sur l'ensemble des pages concernées. Autre technique encore utilisée pour factoriser la production de pages statiques : les SSI (server side includes). Si le serveur web est paramétré pour les accepter, il est possible d'appeler le même fichier HTML à partir de plusieurs autres pages, ce qui permet de réaliser assez facilement des menus de navigation modifiables depuis un seul fichier.

Globalement, cependant, le modèle statique ne s'avère guère souple en terme d'évolutivité. Autre limitation, à mesure que le site se développe : l'équipe de contributeurs s'agrandit, se diversifie, et il devient difficile de demander à chacun de manipuler du HTML.

En revanche, ce modèle permet de très bonnes performances techniques d'accès au site et une bonne résistance aux montées en charge, le serveur « n'ayant qu'à » fournir des fichiers, au contraire du modèle dynamique qui repose sur le calcul de la page.

1.2 La génération dynamique

Du fait de sa vocation première, la présentation d'informations de manière hypertextuelle, le langage HTML a très vite montré ses faiblesses pour tout ce qui concernait l'interaction avec l'internaute. Ce dernier n'a que le choix du lien sur lequel il peut cliquer. Les développeurs ont très rapidement souhaité aller plus loin qu'une simple restitution de pages : recherche dans des bases de données, personnalisation, transactions, etc. Les premières tentatives furent permises par les scripts CGI (common gateway interface) qui jouent le rôle de « passerelles » entre des données d'origines diverses (base de données, formulaire renseigné par l'internaute, etc.) et la page de diffusion que consulte l'internaute sur son écran. Le CGI est toujours bien présent sur le web, que ce soit pour assurer des interactions relativement primaires, comme l'envoi d'un message au webmestre par le biais d'un formulaire, l'animation d'un forum, la gestion des campagnes publicitaires, mais également pour assurer des services fondamentaux tels que la recherche au sein du site. C'est également un moyen de permettre l'accès *via* le web à des bases de données qui ont été conçues avant son développement.

À partir du milieu des années quatre-vingt-dix furent développés plusieurs techniques et langages de programmation pour rendre les sites web dynamiques. Une grande confusion a d'ailleurs régné pendant longtemps autour du terme même de « dynamique ». L'appellation DHTML (Dynamic-HTML) y est pour beaucoup, puisqu'elle a pu faire croire à une évolution du langage HTML, alors qu'elle représente un concept initié au départ séparément par

Microsoft et Netscape pour rassembler les dispositifs permettant d'animer une page web du côté client. Le Javascript et les calques constituent ainsi un moyen de créer des menus déroulants, chose impossible avec des pages uniquement composées de HTML. Mais le Javascript est compris dès l'origine dans la page et c'est le navigateur qui l'interprète, ce qui limite grandement ses utilisations potentielles. Flash, une technologie très répandue permettant les animations vectorisées, nécessite également l'installation d'un lecteur spécifique en conjonction avec le navigateur.

Au contraire des langages de script qui s'exécutent du côté serveur, plus complexes et plus riches à la fois. Microsoft propose en 1996 son langage ASP (Active Server Pages) ; le PHP (Hypertext Preprocessor), élaboré plus tôt dans un cadre strictement personnel, se diffuse sur le web à partir de 1997 ; enfin les JSP (Java Server Pages) fournissent quelques mois plus tard la réponse du monde Java. Ces trois technologies permettent la diffusion dynamique des informations de l'entreprise. Quand le serveur web reçoit une page écrite selon l'un de ces trois langages, il recourt à un module ou à un serveur spécifique (dit d'« application ») pour interpréter le code non HTML. Cette interprétation donne lieu à l'exécution de programmes tiers, à des applications web, à l'échange de certaines données, pour finalement retourner au serveur web une page lisible par le navigateur de l'utilisateur. Voilà pour la logique d'ensemble. Dans le détail, la différenciation entre les trois standards fait l'objet d'âpres discussions entre développeurs. Mais c'est en amont que le véritable enjeu se situe désormais autour de la plate-forme d'intégration des différents serveurs. Le domaine du *middleware* est en effet le terrain d'une vive concurrence entre J2EE, l'environnement développé par la communauté Java, et .Net, la riposte de Microsoft.

On notera que les deux approches du dynamisme (côté client et côté serveur) ne sont pas concurrentes mais complémentaires : le Javascript permet par exemple de contrôler depuis le client la cohérence des données qu'un internaute envoie par formulaire au sein d'un système de gestion de contenu.

Nous sommes donc passés en quelques années d'une architecture à deux niveaux dans le cas du Web statique (le serveur répond au client) à une architecture à trois niveaux (appelée plus communément « trois tiers ») distinguant l'interface, le traitement et la gestion : le client (le navigateur s'occupe de l'interface) demande une ressource à un serveur d'applications (traitement) qui se charge de calculer la réponse après avoir consulté un serveur secondaire (qui gère les données). Au cœur du système se trouve le serveur d'application, car c'est lui qui prend en charge l'ensemble des fonctionnalités permettant aux postes clients d'utiliser ensemble les applicatifs (gestion des sessions et des montées en charge, accès aux sources de données).

C'est globalement sur ce modèle que sont conçus tous les systèmes de gestion de contenu actuels avec, *a minima*, un SGBD-R (système de gestion de bases de données relationnelles) du type d'Oracle ou de MySQL. On comprend néanmoins que le développement web est désormais une activité intégrée à un ensemble plus vaste : le choix entre Microsoft et Java comme plate-forme d'intégration, si tant est qu'il soit si tranché, ressortit plus aux directions informatiques qu'aux services opérationnels.

L'interconnexion de plusieurs sources de données et leur mise en relation avec les informations fournies par l'internaute ouvre la voie à l'interactivité, la personnalisation et le transactionnel (dernier point fondamental pour le développement du commerce électronique). En contrepartie, un tel modèle se fondant sur le calcul systématique des pages implique, en théorie, des temps de restitution bien supérieurs à ceux du modèle statique.

◆ 2. La gestion de contenu collaborative

Cela fait maintenant plusieurs années que la gestion de l'information, et donc sa diffusion, ne sont plus du ressort des seuls informaticiens. Les communiquants ont repris leurs droits, parfois même de manière excessive. La « gestion de contenu web » (*web content management*) dispose de ses propres méthodes et outils afin de permettre à un groupe hétérogène de personnes de participer aisément et collectivement à la réalisation d'un site Internet.

2.1 À la croisée de plusieurs filières technologiques

La gestion de contenu est un formidable concept à géométrie variable : fondamentalement, c'est une préoccupation qui, depuis plusieurs années, réunit de nombreuses professions. Mais le concept en tant que tel connaît une grande fortune depuis les années quatre-vingt-dix. L'observateur qui jetterait un regard rapide dans le rétroviseur, sans trop se soucier de la chronologie précise des dates, pourrait ainsi se remémorer quelques grandes phases de développement des technologies de l'information : la démocratisation de l'Internet avec l'arrivée du web, sa déclinaison en intranet, la fièvre du portail, qu'il soit d'entreprise ou grand public (souvenons-nous du défunt Vizzavi.fr), puis enfin la gestion de contenu.

Cette dernière apparaît à la confluence de paradigmes technologiques en développement depuis de nombreuses années. La Ged (gestion électronique des documents), est apparue au fil des années quatre-vingt, notamment avec le développement de la capacité des supports de stockage ; ont émergé avec elle des problématiques de formats, de suivi de versions ou encore de partage

de documents électroniques, toujours discutées aujourd'hui. Les spécificités de certains secteurs, comme l'aéronautique ou l'édition juridique, ont permis de travailler parallèlement sur la structuration des documents. La normalisation du langage SGML (Standard Generalized Markup Language) en 1986 marque ainsi une grande date, puisque son principe fondateur de séparation stricte entre le fond et la forme est repris par le langage XML (eXtensible Markup Language), actuellement au cœur de la problématique de gestion des contenus.

Le *groupware* est un autre concept en gestation depuis de très nombreuses années : dès les années soixante-dix, des chercheurs élaborent des méthodes pour faire mieux travailler ensemble les membres d'une même équipe ou d'un même projet[2]. Différentes applications en sont issues, telles que la messagerie (synchrone ou asynchrone), les forums, la téléconférence, l'agenda partagé ou encore les espaces de travail partagés. En outre, on associe fréquemment le *groupware* au *workflow* puisque celui-ci permet de modéliser et gérer l'ensemble des tâches et des acteurs impliqués dans un processus métier.

Secouez tout cela grâce à la déferlante de l'Internet et vous obtenez le concept global de « gestion de contenus ». En ce sens, la gestion de contenu est un concept daté, puisque c'est une technologie de convergence qui n'aurait pu naître sans les développements évoqués plus haut.

2.2 Un marché pour l'entreprise en forte consolidation

Les « grandes manœuvres », observées sur le marché des solutions destinées aux entreprises depuis deux ans, confirment d'ailleurs l'étendue fonctionnelle du *content management*. On a coutume de distinguer globalement les acteurs par leur origine. Certains proviennent du secteur de la Ged, comme Documentum ou Filenet, tandis que de nouveaux entrants se sont positionnés directement sur la diffusion web, leur milieu de naissance (Interwoven, Tridion, Instranet…). Sans oublier, entre les deux, des acteurs plus généralistes comme Microsoft ou IBM, pour lesquels ce marché ne saurait être négligé et qui ont les moyens de leurs ambitions : être présents depuis le *back-office* jusqu'au *front-office* de l'utilisateur[3]. Les acquisitions récentes témoignent du souci des différents acteurs de répondre le plus largement au spectre fonctionnel de la gestion de contenu. Ainsi, plusieurs éditeurs de solutions de *content management* ont acquis des sociétés spécialisées dans le collaboratif

2 « GROUPWARE is intentional GROUP processes and procedures to achieve specific purposes plus softWARE tools designed to support and facilitate the group's work », Peter et Trudy Johnson-Lenz (New Jersey Institute of Technology, 1978)

3 Le lecteur pourra consulter en annexe un tableau des principales solutions sur le marché de la gestion de contenu.

en 2003[4]. Le rachat en mars 2004 de l'éditeur Tower Technology dote les solutions de Vignette de fonctions de Ged et de *records management* supplémentaires.

À côté de ces offres complètes et relativement coûteuses, destinées à de grands comptes, s'est développé un marché libre de la gestion de contenu. Outre le fait qu'elles sont presque toutes disponibles gratuitement[5], ces solutions ont pour point commun de se focaliser sur la diffusion web. Au contraire des solutions d'ECM (*enterprise content management*) précédemment évoquées, elles ne se destinent pas à gérer tous les contenus d'une entreprise, sur tous les supports de diffusion. Spip est un des représentants les plus utilisés en France de ces solutions s'appuyant sur le couple MySQL/PHP pour constituer un site dynamique collaboratif. Son éventail fonctionnel relativement large a déjà séduit nombre d'entreprises, de journaux, mais également de sites publics[6]. Facile à installer et à maintenir, Spip est également un très bon moyen de s'initier aux concepts fondamentaux de la gestion de contenu.

2.3 Les éléments d'un système de gestion de contenu

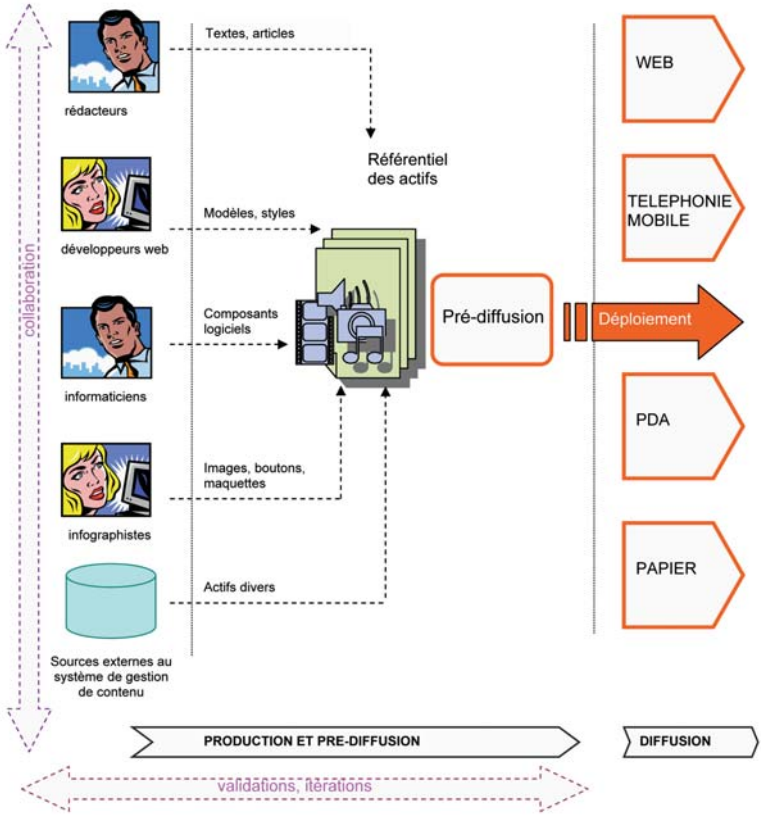
La figure de la page suivante présente les différents éléments qui composent un système de gestion de contenu et les relations qui s'établissent entre eux.

2.4 Les grandes fonctions assurées

2.4.1 Fédérer et gérer l'ensemble des actifs au sein d'un référentiel unique

Le référentiel est au cœur du CMS (*content management system*, système de gestion de contenu). Il rassemble tous les actifs (*assets*) nécessaires à la production des contenus. La notion d'actif est fondamentale dans le concept de gestion de contenu. En effet, au lieu de se focaliser sur un type de fichier à produire, tel qu'on le fait dans un contexte bureautique, le référentiel stocke ici l'ensemble des documents structurés (en XML ou issus de systèmes de gestion de bases de données), des documents non structurés (HTML, images, sons, etc.), mais également tous les composants logiciels destinés à donner de l'« intelligence » aux contenus. La connotation financière du terme actif n'est alors pas fortuite, puisque ce référentiel constitue la richesse informationnelle de l'entreprise. Il faut bien comprendre que, paradoxalement, les actifs sont

4 Par exemple, Vignette a acquis Intraspect et Interwoven l'éditeur I-manage.
5 Rappelons que la gratuité n'est pas la caractéristique fondamentale du logiciel libre.
6 La Documentation française envisage actuellement de porter l'un de ses sites sous Spip et le SIG (Service d'information du gouvernement) a développé pour ses propres sites la plate-forme Agora, à partir de Spip.



nativement passifs[7]. Un site web ne peut exister que grâce à des logiciels qui vont restituer ces actifs ; les images et les pages HTML sont des actifs qui sont activés par le serveur web, à la demande de l'internaute. Quelle que soit la technologie employée pour restituer ces actifs (statique ou dynamique), il est impératif de mettre en place une organisation pour leur gestion.

Si ce sont les actifs dérivés qui seront finalement diffusés, il faut néanmoins en conserver une version source : par exemple, une version Photoshop de toutes les images JPG présentes sur le site, afin de pouvoir les retravailler ultérieurement. Inversement, il n'est pas toujours indispensable de conserver tous les actifs dérivés.

7 Gestion de contenu web : une approche collaborative / Russel Nakano. – Paris : Vuibert, 2002. – 229 p. (Entreprendre Informatique)

Le référentiel a donc pour mission d'assurer la cohérence des actifs (notamment la gestion des versions), leur stockage, leur indexation et leur restitution.

Évidemment, la conception d'un référentiel unique paraîtra illusoire et même dangereuse à certains. La plupart des systèmes s'efforceront plutôt de relier les différentes sources d'information grâce à des connecteurs. Cependant, les solutions émergentes de bases de données XML natives (TextML d'Ixiasoft ou Xyleme Zone Server) permettent de constituer des entrepôts de données non structurées.

La maintenance d'un tel système ne saurait se faire sans de solides arguments… documentaires, hésite-t-on presque à écrire tant le terme paraît décalé. Et c'est vrai que l'on peut s'interroger sur la pertinence d'un terme qui contient en lui-même un objet, le document, désormais complètement éclaté entre ses diverses composantes.

Il n'en reste pas moins qu'une telle hétérogénéité impose d'autant plus le recours aux métadonnées, ces fameuses « données sur les données » qui permettent de les caractériser[8]. Elles concernent aujourd'hui tous les types de contenus : par exemple, de nombreuses recherches sont menées pour indexer automatiquement les flux vidéo. Elles représentent une valeur ajoutée indéniable par rapport à la stricte indexation en texte intégral. Il devient possible d'avoir recours à une véritable taxonomie métier (modernisation de notre bien connu thésaurus) et d'optimiser la recherche de contenus. Mais au contraire de nos systèmes purement documentaires, les métadonnées peuvent représenter plus que les concepts, notamment pour suivre la vie du document ou bien encore pour gérer les droits afférents à sa diffusion. Il pourra être utile alors de se conformer à une ontologie existante, telle que celle du modèle Dublin Core[9], pour favoriser l'échange de ces données avec d'autres systèmes. Dans ce domaine, les solutions d'origine documentaire s'avéreront peut-être plus complètes fonctionnellement.

Le moteur de recherche demeure un des fondements du référentiel de contenus. Il permet de retrouver les contenus, aussi bien en interne pour les contributeurs qu'en externe pour les internautes. Pourtant, l'indexation se révèle le plus souvent relativement limitée. De nombreux sites se dotent ainsi d'un moteur spécifique en diffusion qui, au lieu d'indexer les contenus dans le référentiel originel, indexe les pages obtenues en diffusion. Mais, pour cela, il faut veiller à ce qu'un point d'entrée, pas forcément visible du public, permette de parcourir toute l'arborescence des contenus de lien en lien.

8 Lire à ce sujet : Les métadonnées : accès aux ressources électroniques / Marie-Elise Fréon, *in* : La recherche d'information sur les réseaux : cours Inria, 30 septembre – 4 octobre 2002, Le Bono. – Paris : ADBS Editions, 2002.
9 http://dublincore.org

Plus le système de gestion de contenu a vocation universelle dans l'entreprise, tant en termes d'actifs gérés que de supports de publication, moins l'organisation interne des contenus est calquée sur une publication particulière. En revanche, s'il sert à produire un unique portail sur le web, l'administrateur aura soin de respecter la plus grande cohérence entre l'interne et l'externe. La structure d'un site web statique correspond à celle d'un système de fichiers traditionnel, elle est hiérarchique : toutes les rubriques partent d'une racine unique. L'arborescence étant un modèle très prégnant en informatique, c'est le plus simple à gérer.

Néanmoins, le besoin apparaît parfois de casser ce modèle hiérarchique pour fonctionner de manière plus relationnelle. Même si les « lianes » du modèle hypertextuel permettent d'accéder transversalement à toutes les pages d'un site, depuis n'importe quelle autre page, la page accédée n'en demeure pas moins sur la branche de l'arborescence sur laquelle son concepteur l'a déposée. Le recours à une liane peut s'avérer perturbant dans le cadre de la navigation de l'internaute. Un même contenu doit quelquefois être diffusé dans plusieurs rubriques du site ; par exemple, une même fiche produit se retrouvera dans le catalogue commercial et, pour une courte durée, dans la zone des promotions. On remarquera d'ailleurs la transition lexicale entre « pages » et « contenus ». Attention, tous les systèmes ne gèrent pas aussi facilement le multi-rattachement ; c'est notamment une des limites gênantes de Spip.

La conception de son plan d'organisation de contenus, que l'on ait choisi le modèle statique ou dynamique, est donc une affaire importante, dans laquelle les documentalistes sont amenés à faire valoir leurs compétences. La problématique principale étant rien de moins que de prévoir le futur : à savoir, concevoir une structure qui, dans ses grandes lignes, n'évoluera guère au fil des années. Pour cela, il est indispensable de bien catégoriser le fonds actuel afin d'en conceptualiser les principales caractéristiques.

La dénomination des actifs est enfin très importante. L'absence de plan en la matière est souvent flagrante sur de nombreux sites statiques : les images et les pages ont des noms de fichiers très hétérogènes, sans aucune justification morphologique. Pourquoi aller jusque-là dans l'harmonisation ? D'abord, pour faciliter la gestion de ses actifs et donc leur réutilisation ultérieure. Mais des pages clairement nommées et organisées dans une arborescence simple renforcent également la compréhension de la navigation par l'internaute et le référencement du site par les moteurs de recherche externes. Sont à proscrire tous les caractères interdits ou mal interprétés (espaces, accents, etc.) et les dénominations ésotériques (préférer « catalogue » à « catprod04 » comme nom de répertoire).

On le constate néanmoins, les systèmes de gestion de contenu génèrent fréquemment des URL très exotiques[10], composées notamment de signes ? et & pour inclure des paramètres. Des techniques, regroupées sous le vocable d'*URL Rewriting*[11] existent pour convertir les adresses de ce type en une succession de répertoires contenant finalement un fichier HTML.

2.4.2 Gérer distinctement le fond et les formes

Aujourd'hui encore, la majeure partie des pages HTML parcourues sur le web mêlent allègrement éléments de fond et attributs de forme. Le texte est encadré de balises qui permettent de définir le style de la police (*font type*), sa taille (*font size*) ou bien encore sa couleur (*font color*). Grâce à un éditeur HTML, le texte d'une page est aussi simple à mettre en forme qu'avec son traitement de texte habituel. Résultat : dès que l'on veut modifier le style des titres des pages, il faut les modifier sur chacune d'elles.

L'un des principes forts de la gestion de contenu est la séparation stricte de la forme et du fond. Pourquoi ? Parce que l'une et l'autre relèvent de compétences et de métiers différents et qu'ils doivent pouvoir évoluer indépendamment. Disposer d'une maquette pour l'ensemble du site renforce en outre son homogénéité formelle, et déconnecter le style du contenu permet de doter ce dernier des apparences les plus adaptées aux différents médias. Enfin et surtout, cette séparation permet d'envisager sereinement l'avenir en se dotant de la capacité d'adapter le contenu à tout nouveau dispositif de lecture encore inconnu lors de sa conception.

Un tel mode de fonctionnement implique de fortement structurer l'entité que l'on se destine à publier. En la matière, le modèle dominant est celui de l'article de presse, particulièrement dans les systèmes de publication orientés web. Un modèle qui s'explique notamment par les motivations qui ont conduit leur développement : Spip a été conçu pour produire Uzine, le pionnier des *webzines* alternatifs, et Cofax a été conçu par le groupe de presse Knight-Ridder pour ses journaux en ligne.

On distingue deux types de structuration d'un document. La structuration fonctionnelle, la plus répandue, décompose l'objet en blocs assurant chacun une fonction différente : titre, auteur, date, résumé, corps, etc. La structuration sémantique découpe quant à elle le document en éléments ayant des significations spécifiques dans le contexte de la publication : réalisateur, acteurs, producteurs sont des champs de structuration adaptés à un site de cinéphiles. Plus les contenus seront finement structurés, plus grandes seront

10 Un exemple parmi d'autres, pris sur le site de *Libération* : www.liberation.fr/page.php?Article=207157
11 www.webrankinfo.com/analyses/autres/url-rewriting-debutants.php

les possibilités de modulation et de personnalisation à l'affichage. Il est judicieux de disposer de plusieurs modèles de documents, pour s'adapter au mieux à la production éditoriale, mais pas trop non plus pour ne pas dérouter les contributeurs et conserver une certaine homogénéité à son fonds. Selon les cas, la latitude pour créer de nouveaux types de contenus peut s'avérer un critère décisif quant au choix de son CMS.

En regard de ces modèles structurels se trouvent les modèles formels, appelés également gabarits ou *templates*. Leur fonction est double : sélectionner les contenus qui seront publiés et indiquer sous quelle forme. On pourra décider ainsi que la page d'accueil d'une rubrique ne comprend que le titre, l'auteur, la date et le résumé des dix derniers articles, classés chronologiquement. Ce besoin se traduit sous la forme d'une requête spécifique adressée au référentiel de contenu.

Pour la mise en forme, la solution la plus simple est d'avoir recours à des pages HTML contenant en leur sein des balises spécifiques au CMS et permettant de déclarer les requêtes. Si les contenus du référentiel sont au format XML, il est alors préférable de les mettre en forme grâce à une « transformation » XSL (XSLT, eXtensible StyleSheet Language Transformation). Un « processeur » s'appuie sur la feuille de style XSL pour générer un document lisible dans un navigateur web. C'est la technologie retenue pour la diffusion des pages du guide *Vos droits et démarches* disponible sur le site Service-public.fr.

2.4.3 Favoriser la collaboration en l'organisant

C'est explicite dans le schéma de la précédente partie : la valeur ajoutée d'un système de gestion de contenu se révèle d'autant mieux que les contributeurs sont nombreux. Ces derniers proviennent généralement d'horizons très divers. Les informaticiens écrivent les applications spécifiques pour animer les actifs, les développeurs HTML conçoivent les modèles de présentation des pages, les infographistes réalisent la couche graphique de l'interface et, enfin, les rédacteurs produisent les contenus. Parmi ces derniers, certains écrivent, d'autres relisent et enfin un petit groupe d'individus, quand il ne s'agit pas du seul responsable éditorial, valide la mise en ligne des contenus. Les droits affectés à ces différentes personnes varient en fonction de l'organisation des contenus et des missions des contributeurs (proposer, consulter, mettre à jour, valider, publier, etc.). Par exemple, le correspondant éditorial du service marketing disposera d'un droit d'écriture uniquement sur la rubrique « Opérations », mais sans possibilité de toucher à sa structure.

La publication de contenus est un processus relativement balisé qui peut être modélisé sous forme de tâches et de rôles. Tous les CMS proposent donc des

fonctions de *workflow* plus ou moins étendues, selon le niveau de sophistication du logiciel. Les mauvaises langues caricaturent fréquemment en assimilant ce concept à une tentative brutale de taylorisation de l'immatériel. Il est vrai que mal préparée, la mise en place d'un *workflow* peut conduire à certaines aberrations, particulièrement quand elle concerne des processus intellectuels. Mais bien conçu, il peut se calquer sur une chaîne de publication pré-existante pour organiser les tâches, permettre leur suivi, tout en conservant une certaine souplesse dans l'articulation entre les différentes fonctions.

Russel Nakano[12] distingue plusieurs éléments fondamentaux dans un *workflow* : la personne, le processus, la tâche, l'interaction et la notification. Un contributeur est « notifié » par courrier électronique ou dans l'interface du système des « tâches » qui lui sont affectées ; c'est en « interaction » avec les autres contributeurs qu'il va les mener à bien. Selon l'auteur américain, le temps relatif à une tâche particulière peut se diviser en temps de réflexion (action), temps d'attente occupé et temps mort. Plusieurs processus parcourant simultanément la chaîne éditoriale, un même contributeur peut être actif sur un processus et donc en attente sur un autre. Le but du *workflow* est alors de faire coïncider au mieux ces différents temps pour minimiser les zones de temps morts. Tout en respectant les conditions de travail du contributeur, bien entendu !

La collaboration entre de nombreux contributeurs travaillant simultanément sur certains documents peut créer ponctuellement des conflits de versions. Le CMS doit assurer deux fonctions directement héritées de l'univers de la Ged : le *check-in check-out* et la gestion des versions. Le premier permet de bloquer un fichier lorsqu'il est ouvert par l'un des contributeurs ; les autres personnes qui souhaitent y accéder ne peuvent que l'ouvrir mais sont informées du nom de la personne qui le « bloque » et peuvent donc la contacter. La gestion des versions consiste à revenir en arrière d'un à plusieurs niveaux ; cette gestion porte sur des actifs individuels mais peut également concerner une édition globale du site.

Nakano distingue dans son ouvrage cinq opérations signifiantes dans un système fortement collaboratif.

• *La soumission* d'un actif marque provisoirement la fin du travail du contributeur, au sens large. Elle consiste en la copie de l'actif d'une zone de travail vers la zone de pré-diffusion en incluant l'identité de la personne qui soumet, l'heure de soumission ainsi qu'un commentaire. Les actifs de la pré-diffusion ne peuvent être modifiés, ils ne peuvent qu'être effacés ou remplacés par une

12 *Ibid.*

nouvelle version provenant d'une zone de travail à la suite d'une itération enclenchée par la révision de l'actif par un autre membre de la chaîne éditoriale.

• *La comparaison* identifie les actifs nouveaux, modifiés et supprimés dans la zone de pré-diffusion par rapport à une zone de travail donnée. Les actions déclenchées suite à cette comparaison permettent d'assurer la cohérence des différentes zones de travail.

• *La mise à jour* consiste à copier les modifications depuis la pré-diffusion vers une zone de travail donnée. Plus cette opération est fréquente, moins la zone de travail est désynchronisée.

• *La fusion* résout les conflits entre une zone de travail et la pré-diffusion. Elle prend en compte des modifications qui nécessitent plus qu'une simple opération de copie.

• Enfin, *la publication* crée sur le ou les serveurs de diffusion une copie en lecture seule de la pré-diffusion complète, enregistre l'identité de celui qui publie, l'heure de publication ainsi qu'un commentaire.

Nous venons de le voir : chacune de ces opérations se traduit concrètement en mouvements d'actifs entre zones. La dimension sociale est bien présente, puisque seuls certains contributeurs, en fonction de leur place dans l'environnement collaboratif, peuvent déclencher les opérations les plus cruciales comme la publication.

2.4.4 Suivre et contrôler la chaîne éditoriale, des développements aux déploiements

Le contributeur, qu'il soit rédacteur, développeur web ou bien infographiste, est clairement identifié selon ses droits et les tâches qu'il a à accomplir. Le système accompagne les multiples intervenants tout au long de la chaîne éditoriale. Cette chaîne se décompose en plusieurs cycles, de la création de l'actif jusqu'à la publication d'une version.

Le premier cycle consiste à *développer* les actifs sur sa propre zone de travail. Le développeur web, par exemple, crée les modèles de pages à partir de son éditeur. Une des spécificités de ce type de développement est la rétroaction directe. Le développeur écrit du code, le teste, le corrige, le teste, etc. Il continue ainsi de manière itérative jusqu'à parvenir à une version de soumission.

Intercalées dans ce premier cycle, les phases de *comparaison* et de *mise à jour* éliminent les différences existant entre la zone de travail du contributeur et la zone de pré-diffusion. Dans tous les cas, c'est cette dernière qui joue le rôle de référence en cas de conflit.

Le cycle de *révision* consiste ensuite à soumettre, comme nous l'avons vu dans la partie précédente, ses actifs à un vérificateur (rédacteur en chef, directeur artistique, etc.).

Enfin, dans l'idéal (!), un cycle de *test principal* confie à un responsable qualité la tâche de tester les actifs selon des critères ergonomiques et d'accessibilité.

Ces quatre cycles se succèdent rapidement dans le meilleur des cas, mais ils peuvent bien évidemment boucler si des problèmes se présentent. Dans tous les cas, le CMS se doit de tracer l'état des différents actifs tout au long de la chaîne éditoriale, de détecter les engorgements éventuels et d'enregistrer des données de suivi.

Une fois les actifs révisés et approuvés, nous entrons dans la phase de déploiement. Cette dernière est bien évidemment d'autant plus délicate que le CMS gère plusieurs sites, plusieurs supports et sur des échelles de taille importantes. Chaque déploiement représente une version du site, à partir de laquelle le système de gestion de contenu doit permettre un retour en arrière vers une version antérieure (*rollback*). Le déploiement ne concerne que les modifications entre la pré-diffusion et les différents serveurs de diffusion, ces modifications sont d'ordre positif (ajout ou modification d'un actif sur la diffusion) ou négatif (suppression d'un actif sur la diffusion). Comme les actifs sont le plus souvent liés entre eux (une page HTML fait appel à de nombreuses images), et qu'un dysfonctionnement qui interromprait brutalement le déploiement pourrait avoir de graves conséquences, certains CMS permettent un « déploiement transactionnel » : les modifications n'apparaissent qu'une fois qu'elles sont totalement transférées, sinon c'est le processus global qui est à recommencer.

Transversalement à ces différents cycles, somme toute relativement linéaires, de la création au déploiement, les documents ont leur propre vie que le CMS se doit de gérer au mieux. Un article peut ainsi être rédigé bien avant sa date de publication prévue et demeurer dans le référentiel jusqu'à ce que le cycle de déploiement automatique lui permette d'exister en ligne. Si l'article dispose d'une durée de vie prédéfinie, le système le retire automatiquement de l'espace de diffusion à la date d'expiration ; s'il n'en dispose pas, la tâche de nettoyer périodiquement le site revient à un administrateur humain. Quelle vie après la diffusion ? L'article est-il archivé ou bien détruit immédiatement ? Le rédacteur qui est à l'origine du document doit-il suivre par notification toutes ces étapes ? D'autant plus qu'un article peut avoir une existence en ligne bien plus complexe : quelques jours en page d'accueil, un mois dans un dossier particulier, puis trois ans dans les archives payantes.

On comprend donc que le déploiement des actifs peut s'avérer parfois d'une réelle complexité. D'autant plus si le site a recours à de la personnalisation. Dans ce cas, les pages s'affichent différemment selon le profil de l'internaute inscrit. Ce qui suppose la connexion du CMS à une autre base d'individus, outre celle des contributeurs : celle des abonnés.

Enfin, nous terminerons cette partie en abordant le déploiement « hors les murs » avec la syndication[13]. Ce concept est actuellement en plein essor, notamment grâce aux CMS orientés web qui en ont grandement simplifié la mise en place. Celle-ci consiste à mettre à la disposition d'autres sites un fichier signalant ses nouveaux articles. Ce fichier, au format RSS (Rich Site Summary), conforme au standard XML, est toujours accessible à la même URL. Le site qui souhaite relayer le flux RSS doit installer au sein de ses pages un lecteur de RSS qui traduira en HTML les balises du fichier. Si l'internaute clique sur un lien, il consulte l'article directement sur le site qui l'a produit. De nombreux éditeurs proposent aujourd'hui gratuitement leur fil de syndication[14].

2.4.5 La problématique encore émergente de l'archivage des pages

L'archivage du World Wide Web, en tant que réseau d'informations, est une problématique qui a désormais droit de cité dans des univers tels que la recherche universitaire[15], la conservation patrimoniale ou encore la production législative. Pour preuve, le « projet de loi sur le droit d'auteur et les droits voisins dans la société de l'information », présenté au Conseil des ministres en novembre 2003, organise le dépôt légal des pages Internet auprès de la Bibliothèque nationale de France et de l'Institut national de l'audiovisuel[16]. Onze bibliothèques nationales et l'association à but non lucratif Internet Archive[17] sont engagées jusqu'en juillet 2006 au sein de l'International Internet Preservation Consortium afin de développer des outils libres appropriés à l'archivage du web.

Au cœur des entreprises, cependant, le site web est encore très souvent un produit émergent pour lequel on peine déjà à consacrer les ressources suffisantes à son développement. L'idée d'un archivage spécifique est encore très peu répandue – et donc peu appliquée. Et que faut-il archiver ? Alors que le projet de loi sus-mentionné pointe l'unité documentaire du site, les entreprises semblent plus préoccupées par la dématérialisation et la traçabilité de

13 Voir aussi le chapitre 1, pages 30-46.
14 *Libération* propose son fil d'actualités à cette adresse : www.liberation.fr/rss.php
15 Lire à ce propos la thèse de doctorat de Mehdi Gharsallah, Archivage du web français : dépôt légal des publications électroniques et préservation patrimoniale. – Saint-Denis : Université Paris 8, 2002.
16 www.culture.gouv.fr/culture/actualites/communiq/aillagon/droitdauteur1103.html
17 www.archive.org

leur information, quels que soient les supports de diffusion. Particulièrement outre-Atlantique, depuis la déflagration Enron qui a révélé à ceux qui l'ignoraient la dimension juridique de l'information. Pour les grands éditeurs de solutions de gestion de contenu, c'est bien le contrôle sur les données de l'entreprise, particulièrement lorsqu'elles sont sensibles ou à l'inverse largement communiquées, qui semble primer sur l'archivage du site en tant que tel. Si l'offre *web content management* de Documentum sait gérer les versions de sites, les responsables de la société préfèrent mettre en avant leur module de *records management*, bien plus en phase, selon eux, avec des impératifs juridiques de plus en plus contraignants pour les entreprises.

Pour ce qui concerne l'archivage de l'entité site, il est bien évident qu'un simple *back-up* des fichiers du site n'est pas suffisant. L'objectif est de figer le site à un instant donné, afin d'en permettre la consultation ultérieure, de manière autonome.

Julien Masanès (BnF) distingue deux méthodes d'archivage des sites[18]. L'archivage côté serveur consiste en une copie de l'ensemble de l'arborescence du site ; on maîtrise ainsi l'ensemble des contenus, même s'ils ne sont pas directement liés (cas des anciennes pages accessibles uniquement au travers d'un moteur de recherche). L'inconvénient est que, hors du HTML, les scripts et les bases de données demeurent dépendants de leur environnement d'exécution. Pour ces dernières, la BnF préconise une exportation des données en XML, pour casser la mise en forme propriétaire, conserver la structure native de la base et faciliter la reprise ultérieure des données.

L'autre méthode consiste à déclencher l'archivage depuis le client, pour obtenir un « cliché » du site tel que pourrait le consulter un internaute le jour de l'archivage. Cette méthode ne résout évidemment pas le problème des « pages profondes » accessibles uniquement au travers d'une zone de recherche ou après saisie d'un mot de passe. En revanche, la réalisation de la version peut s'avérer plus souple grâce à l'un des logiciels disponibles sur le marché. HTTrack website copier[19] est un logiciel libre français (!) relativement performant : il réorganise la structure des liens en relatif (et non plus par rapport à la racine du site web), gère les problèmes liés au renommage des fichiers (caractères interdits, doublons suite à la casse des noms, etc.) et permet éventuellement les mises à jour incrémentales d'archives.

Il est peu pertinent d'envisager de conserver l'aspect dynamique de certains sites : les pages statiques se prêtent mieux à la conservation et le caractère

18 Julien Masanès. – Les problèmes particuliers de la préservation d'Internet, *in* Internet : la mémoire courte ?, séminaire Aristote, 22 avril 2004.
19 www.httrack.com

définitif de la version d'archive ôte tout intérêt aux possibilités de mise à jour des systèmes de gestion de contenu. Certains sites décident ainsi de convertir leurs bases de données en pages statiques au moment de la création de l'archive. C'est cette stratégie qu'a adoptée le Service d'information du gouvernement pour l'archivage des différentes versions du site du Premier ministre. Les pages générées dynamiquement par Coldfusion à partir d'une base de données SQL ont été converties en HTML statique et mises à disposition du public dans quatre répertoires distincts correspondant aux gouvernements successifs[20]. Ce choix remarquable de permettre la consultation des versions de son site est bien entendu lié à la nature toute particulière du site en question. Rares sont les sites qui suivent aujourd'hui cet exemple, même si l'on peut présager tout l'intérêt potentiel pour les entreprises de communiquer sur leur patrimoine hypertextuel.

◆ 3. Les différents types d'accessibilité au service de tous les internautes

3.1 Les performances : dépasser la fracture numérique

Avec la diffusion rapide de l'ADSL (Asynchronous digital subscriber line) dans nos contrées, il devient difficile pour certains développeurs web, qui surfent à grande vitesse au travail et à la maison, de ne pas oublier que les internautes ne bénéficient pas encore tous du haut débit. Pour le premier trimestre 2004, l'Observatoire des équipements multimédias de Médiamétrie dénombre 7 303 000 foyers connectés à l'Internet dont 2 998 000 à haut débit[21] (environ 41 %). Ce qui veut dire que plus de la moitié des Français naviguent encore à bas débit depuis leur domicile. Une considération à prendre en compte, particulièrement si le site que l'on gère a une vocation commerciale ou de service public. D'autant plus que, paradoxalement, la majorité des terminaux mobiles en usage actuellement (téléphones, assistants numériques, etc.) ne sont pas équipés de modems à haut débit.

Au-delà de l'exigence élémentaire de disponibilité de ses propres serveurs, il devient primordial de mesurer la qualité de service telle qu'elle est perçue par les internautes. Des services ASP (*application service provider*) tels que Witbe ou IP-Label mesurent les performances des sites de leurs clients en les consultant régulièrement (d'une fois toutes les heures à toutes les cinq minutes) depuis plusieurs points de France ou du monde. La mesure se fait

20 www.archives.premier-ministre.gouv.fr
21 Les baromètres multimédia : www.mediametrie.fr/web/resultats/barometre/resultats.php?id=996

en bout de chaîne, elle ne pointe donc pas seulement les dysfonctionnements pouvant survenir sur les sites eux-mêmes. Elle porte généralement sur la page d'accueil, mais, selon les statistiques de consultation (voir la section suivante), d'autres pages peuvent s'avérer tout aussi cruciales à surveiller.

Ces services permettent également de surveiller des cheminements bien particuliers, constitués de plusieurs pages et même de formulaires à remplir, quels que soient les serveurs sur lesquels se trouvent les pages. Il est ainsi possible de s'assurer du bon déroulement d'une transaction commerciale, même si celle-ci comporte plusieurs applicatifs distincts (catalogue puis système de paiement). Le télédéchargement de chaque objet composant une page auditée est lui-même mesuré, ce qui peut se révéler tout à fait précieux pour optimiser la conception d'une page.

3.2 Le code : appréhender les différents dispositifs de lecture

Il y a quelques années de cela, la guerre que se menaient les sociétés Netscape et Microsoft se traduisait concrètement par des atteintes régulières au standard HTML, tel qu'il est décrit par le W3C (World Wide Web Consortium). En terme de production de sites, les développeurs bien intentionnés en vinrent à pratiquer le développement *cross-browsers* : un script sur la page d'accueil du site permet d'aiguiller l'internaute vers les pages spécifiquement adaptées à son navigateur. La guerre entre navigateurs est derrière nous, mais les dangers de retour à un web propriétaire n'en ont pas pour autant disparu. En effet, par un de ces cercles vicieux dont les technologies de l'information ont parfois le secret, les développeurs web travaillent prioritairement pour la plate-forme la plus répandue sur le marché – Microsoft Internet Explorer sous Windows – tandis que les internautes les plus assidus savent que l'usage de cette plate-forme leur garantit un rendu optimum.

Le chef de produit sensibilisé à ce problème de compatibilité se retrouve devant un choix stratégique : privilégier la compatibilité totale (forme et fond) pour tous les navigateurs à partir de certaines versions spécifiques (par exemple, à partir d'Internet Explorer 4 et plus) ou s'assurer de la lisibilité du contenu sur tous les navigateurs, indépendamment de la forme. Le premier choix impose de se conformer au standard HTML 4 sans avoir recours aux feuilles de style normées CSS2 (*Cascading Style Sheets*) pour la mise en forme, puisque ce standard du W3C n'est pleinement lisible, dans la famille Microsoft, qu'à partir du navigateur Internet Explorer 5[22]. Aujourd'hui, il est vrai que les navigateurs inférieurs à cette version ne représentent qu'une part

22 http://dicolive.media-box.net/docCSS/css.php?orderByType=1

négligeable en pourcentage de l'équipement des internautes. Mais lorsque l'on prend les valeurs absolues, les résultats obtenus sur des sites à forte fréquentation peuvent amener à réfléchir : sur Service-public.fr, au mois d'avril 2004, seules 0,4 % des visites ont été réalisées avec un navigateur de type IE 4.x ; en valeur absolue, cela représente tout de même près de 8 500 visites !

Bien entendu, cette hypothèse de travail n'est pas la solution d'avenir ; elle ne peut être qu'un maintien transitoire d'un ancien état de fait, en attendant que les parcs de navigateurs se mettent à jour et deviennent entièrement compatibles avec le CSS2. L'avenir est à la séparation de la forme et du contenu, comme nous l'avons vu précédemment, et les défenseurs des standards font justement remarquer que si les anciens navigateurs ne peuvent pas reproduire fidèlement la mise en forme permise par le CSS2, ils affichent néanmoins correctement le contenu, d'autant plus que celui-ci aura été conçu et rédigé dans un ordre de lecture logique. Et les avantages de l'externalisation de la mise en forme sont nombreux : indépendance vis-à-vis de l'éditeur, de la plate-forme ou encore de l'appareil de lecture ; facilité de maintenance, puisque les attributs de présentation sont centralisés ; optimisation de la bande passante, grâce à la suppression de nombreux artifices de présentation fondés sur des images… Enfin, le recours au CSS2 est un pas de plus vers l'accessibilité de ses pages à tous les dispositifs de lecture disponibles sur le marché.

Les propriétés de positionnement permettent notamment de s'affranchir de la ruse des tableaux invisibles pour présenter le contenu en plusieurs colonnes : après plusieurs mois de mises à jour et d'évolutions non planifiées, l'imbrication de multiples tableaux constitue un véritable casse-tête pour lire la page hors du navigateur cible.

La mise en œuvre des standards HTML 4.01 et CSS2 représente une étape importante avant l'adoption du XHTML (eXtensible HyperText Markup Language). Cette recommandation du W3C joue en effet le rôle de pivot entre l'« ancien monde » du HTML et les nouveaux rivages du XML : structuration, séparation absolue de la forme et du fonds, modularité et évolutivité, etc.

3.3 L'ergonomie : faciliter tous les usages

Travailler à faciliter l'accès des personnes déficientes, c'est contribuer à un meilleur accès pour toutes les populations d'internautes. Cette affirmation universaliste, promue en substance par la plupart des sociétés du secteur de l'accessibilité, est tout à fait en phase avec le paradigme de la séparation de la forme et du contenu. Au-delà de l'aspect « politiquement correct », la démocratisation du web, le vieillissement de la population des internautes et

la diversification des terminaux de consultation font que les conditions de lecture se sont considérablement diversifiées – handicap ou pas.

Depuis plus de cinq ans, le W3C est engagé dans une réflexion sur l'accessibilité des sites au travers de sa *Web Accessibility Initiative* (WAI, Initiative pour l'accessibilité du web). La version 1 de ses recommandations (*guidelines*), sortie en 1999, associe à chacune de ses quatorze directives des « points à contrôler » (*checkpoints*) organisés en trois niveaux de priorité, selon que le développeur « doit », « devrait » ou « peut » satisfaire à ces critères.

En France, le travail du W3C a été adapté, avec des variantes, par l'association BrailleNet au travers du label Accessiweb[23]. Décliné lui aussi en trois niveaux (bronze, argent et or), il repose sur le même principe d'une mise en conformité progressive. Enfin, l'Adae (Agence pour le développement de l'administration électronique) propose un *Référentiel accessibilité des services Internet de l'administration française*, une version déclinée d'Accessiweb complétée par des critères ergonomiques propres aux sites publics (une importance toute particulière est ainsi accordée aux formulaires).

Globalement, une grande partie des critères objectifs peut être assez facilement satisfaite, au prix d'un travail de relecture et de recodage parfois long et astreignant. L'exemple le plus parlant est celui de la balise ALT (commentaire alternatif d'une image) qui doit être systématiquement renseignée. Mais certains critères, plus subjectifs comme la qualité rédactionnelle des textes, sont plus difficiles à évaluer, particulièrement dans une structure où le développement des sites relève d'une gestion de projet transversale à plusieurs équipes. Comme pour les langues, l'immersion totale est un moyen de s'y mettre : le site « Plongez dans l'accessibilité »[24] propose de se consacrer chaque jour à un point particulier de l'accessibilité. Exemple au dix-septième jour : utiliser la balise « *acronym* » pour insérer le développé de chaque abréviation et sigle employé. Sur le plan de l'accompagnement, de plus en plus d'acteurs, associations, prestataires ou sociétés de conseil proposent des missions d'audit et d'optimisation des sites au regard des référentiels évoqués.

Mais une fois atteint un certain niveau d'accessibilité, encore faut-il maintenir cette qualité constante, voire l'améliorer, en sensibilisant les équipes de développement et en consentant un effort important de formation. À moyen terme, le développement d'un site accessible et conforme aux standards est un gage de meilleure évolutivité ; à court terme, il serait malhonnête de ne pas reconnaître qu'il représente un investissement humain spécifique.

23 www.accessiweb.org/fr/Label_Accessibilite/
24 www.la-grange.net/accessibilite

◆ 4. Des fréquentations à surveiller

4.1 Serveur contre marqueurs

Que ce soit pour mesurer le succès de son site personnel ou étudier très finement les usages que les internautes font de son portail, le suivi de la fréquentation est aujourd'hui un aspect indissociable de la production. Nous parlerons ici de fréquentation plutôt que d'audience, ce dernier terme désignant plus l'analyse qualitative de publics et de leurs usages, ce qui n'est pas l'objectif des solutions statistiques évoquées dans cette partie. La démarche évoquée ici est dite *site-centric*, puisqu'elle consiste à mesurer précisément les usages qui sont faits d'un site.

Historiquement, la première technique de mesure de la fréquentation consiste à exploiter les fichiers de *logs* générés par le serveur lors de chaque connexion d'internaute. Chaque clic sur un lien envoie en effet au serveur plusieurs informations mémorisables et exploitables : l'adresse IP, le domaine, le navigateur, le système d'exploitation, la résolution, la page depuis laquelle le clic a été effectué, etc. Une fois consolidées, ces « variables d'environnement » permettent d'obtenir une vision très simple de l'usage de son serveur. On parle communément de *hits* de consultation.

Des logiciels ont été développés pour apporter une analyse plus sophistiquée de ces données de consultation. Ils s'appuient sur deux technologies distinctes. La première consiste à exploiter localement les fichiers de *logs* pour proposer des analyses plus sophistiquées que celles proposées en standard par les logiciels serveurs. On citera dans cette famille Webtrends, ainsi que les solutions libres WebAlizer ou AWStats qui sont très répandues chez les fournisseurs d'accès. Cette technologie a très vite montré certaines limites : c'est une application supplémentaire à installer, les rapports ne sont pas forcément très conviviaux et, plus important, les internautes qui s'arrêtent au cache de leur fournisseur d'accès sont par définition difficiles à comptabiliser (ce qui paradoxalement peut désavantager les sites les plus populaires).

Une autre approche est donc apparue, fondée sur la technologie du marqueur : un lien vers une image transparente, d'un *pixel* sur un *pixel*, est inclus en bas de chaque page que l'on souhaite auditer[25]. Cette image se situe sur le serveur d'un prestataire externe qui comptabilise le nombre de fois que cette image est demandée et depuis quelle page. Avantage de ce regroupement logique des pages, il devient possible de constituer des périmètres

25 Certaines solutions disposent d'une offre gratuite qui, dans ce cas, affiche un logo publicitaire en lieu et place du pixel transparent.

d'analyse plus large qu'un simple serveur : où que soient stockées les pages sur l'Internet, leur consultation est affectée à l'entité indiquée dans le marqueur. La collecte des données, leur exploitation et leur consultation sont assurées par un prestataire externe. Sans se fonder sur des chiffres précis, on peut avancer que cette approche a été adoptée majoritairement par les plus grands éditeurs sur le web. C'est d'ailleurs celle qui est retenue par l'organisme Diffusion Contrôle pour certifier la fréquentation des sites.

4.2 Naissance de la visite

La mesure de la fréquentation s'est heurtée dès sa naissance à un écueil technique : le protocole HTTP (HyperText Transfer Protocol), qui fait circuler l'information sur le web, n'a pas été conçu pour gérer la session d'un utilisateur. Impossible de connaître nativement l'heure à laquelle un internaute particulier s'est connecté sur son site, l'heure à laquelle il en est parti et, surtout, ce qu'il a fait dans l'intervalle.

Les *cookies* sont apparus en 1996 pour pallier cette lacune. Installés sur votre disque dur par un site web particulier, ils permettent d'identifier strictement le micro-ordinateur et par-là même de reconstituer une session de navigation. Suite aux premiers abus autour de la notion de *hits* (certains éditeurs incluaient dans cette notion tous les éléments composant une page, générant ainsi des chiffres de consultation énormes), un consensus a été dégagé entre les différents acteurs de ce secteur naissant pour définir comme « visite » une succession de pages vues consultées par le même poste dans un intervalle de temps entre chaque page inférieur à trente minutes. Si l'internaute s'absente de son poste sur une durée plus longue, une nouvelle visite est ouverte à son retour sur le site[26].

Il faut bien comprendre que, avec la visite, nous quittons le domaine systématique de l'informatique (la page est une unité discrète) pour aborder les rivages de la consolidation et de l'extrapolation. Mais c'est avec le « visiteur unique » que la mesure de fréquentation prend toute sa saveur. Car les *media-planners* ne rêvent que d'une chose : connaître le nombre de visiteurs distincts d'un site. Pour obtenir ce chiffre, les solutions à base de marqueurs s'appuient sur le *cookie* installé en local. Mais comment faire alors pour prendre en compte les quelque 3 % de rebelles[27] qui refusent les *cookies*, ou les maniaques qui les expurgent régulièrement ?

26 On retrouvera les définitions les plus consensuelles sur le site du CESP (Centre d'études des supports de publicité) : www.cesp.fr/docs/fd_docs_pub0002.htm

27 Chiffre avancé par Diffusion Contrôle, l'organisme qui certifie les tirages de la presse et qui s'intéresse désormais également au web ; un chiffre proche de nos propres observations.

Xiti, l'une des solutions leaders sur ce marché de la mesure de fréquentation, utilise une combinaison de plusieurs critères techniques envoyés par le navigateur pour identifier « de manière fiable sur la journée » les visiteurs uniques n'acceptant pas les *cookies*.

Nous obtenons ainsi, par calcul, un nombre de visiteurs uniques sur la journée. Si nous étendons cette période, il est nécessaire d'ajouter une couche de calcul pour dédoublonner les visiteurs. Et, pour corser le tout, on ajoutera qu'il serait plus correct de parler de postes de consultation uniques puisque, dans les lieux publics, plusieurs internautes distincts peuvent se succéder sur le même poste…

4.3 Mieux pistés sur le web que sur le papier

On le comprend donc très vite : nous sommes encore loin de la simplicité des chiffres de tirage de la presse fournis en France par Diffusion Contrôle[28]. Mais si la mesure globale peut encore progresser, les outils spécialisés fournissent aujourd'hui des chiffres qui feraient pâlir d'envie tout patron de presse. Chaque page, chaque rubrique ou groupe de rubriques est audité, un journaliste peut ainsi connaître presque en temps réel[29] la consultation d'un de ses articles. Pour une période donnée, il est possible de mesurer la fréquentation de tout ou partie du site, les principales pages d'entrée ou de sortie ou encore l'équipement utilisé par les internautes (très utile pour les arbitrages en matière d'accessibilité).

Au-delà de ce fondement quantitatif indispensable, les solutions fournissent aujourd'hui des analyses plus poussées à destination des spécialistes du marketing qui recherchent des indicateurs plus qualitatifs. La présence d'un site sur le web peut être étudiée grâce aux indications fournies par les moteurs et annuaires qui ont amené du trafic, ainsi que par les principaux sites pourvoyeurs de visites. On sait combien il est devenu important de tisser les bons partenariats sur la toile et, en la matière, Xiti propose même une fonctionnalité optionnelle qui permet d'isoler la fréquentation provenant de pages internes ou de sites externes. Il devient ainsi possible d'évaluer les retombées d'un échange de liens, d'une lettre de diffusion ou encore de l'achat de mots-clés sur les moteurs de recherche.

Autre analyse fournie : les chemins de navigation les plus fréquents. Selon les solutions, on peut connaître statistiquement les suites de pages les plus empruntées à partir d'une page ou pour accéder à une page. Toujours dans leur effort pour proposer un semblant de sens à partir des chiffres, certaines

28 Quoiqu'en la matière certaines anecdotes viennent nous rappeler régulièrement l'élasticité de ces chiffres.
29 Certaines solutions ne fournissent des chiffres qu'à J+1.

solutions fournissent des critères « d'intérêt » qu'il vaut mieux manipuler avec beaucoup de prudence. Si un « nombre de pages par visite » élevé peut signifier que les internautes ont eu plaisir à naviguer sur le site, il peut également signifier, dans le cas d'un service d'informations pratiques, qu'ils ont bien « patouillé » avant de débusquer l'information recherchée !

Certains pénètrent sans complexe dans le domaine qualitatif et ajoutent à leurs mesures statistiques des données obtenues par panel. C'est le cas de Weborama, dont l'outil Weboscope fournit des informations sur l'âge, le sexe et, évidemment, la catégorie socioprofessionnelle des visiteurs grâce à un « mégapanel » de 300 000 internautes.

Pour pouvoir exploiter ces données sur le long terme, un site important (en nombre de pages) consacrera un temps raisonnable à la conception de son plan de marquage. Toutes les fonctionnalités n'étant pas disponibles à tous les niveaux, le webmestre aura soin de bien doser son effort entre précision potentielle des données et commodité de consultation. Bien que moins décisives car moins visibles des internautes, les décisions en la matière demandent presque autant d'attention que pour l'élaboration de l'arborescence des contenus.

Malgré les tentatives évoquées, il est encore difficile de connaître son internaute « sorti de la masse ». Le site qui voudra étudier la composition sociologique de sa fréquentation ne pourra faire l'économie d'une étude plus particulière, notamment par le biais d'une campagne de sondage en ligne.

◆ 5. Conclusion

Production aisée des pages, optimisation des processus, gestion du référentiel de contenus, suivi des performances et de la fréquentation, la production des sites web ressortit aujourd'hui à plusieurs domaines de compétences : informatique, ergonomie, gestion documentaire, marketing, qualité… L'homme (ou la femme) providentiel(le) n'existe plus, le webmestre a cédé la place aux chefs de projets. La production web est désormais une affaire collaborative qui fait intervenir, de l'amont à l'aval, la plupart des métiers et services de l'entreprise. Mais les métiers existants suffisent-ils ? Nous ne connaissons pas encore l'étendue des bouleversements provoqués par l'évolution des modes de production évoquée dans ce chapitre. Le monde du libre devrait, encore une fois, permettre à de nombreuses petites entreprises de se familiariser avec la gestion de contenus. Et favoriser l'apparition de nouveaux professionnels, dont il est encore trop tôt pour imaginer le nom.

◆ Annexe 1 : Les principales solutions sur le marché de la gestion de contenu web (1)

Broadvision	Broadvision V7.1	http://www.broadvision.fr/
Cofax	Cofax (Content Object Factory)	http://www.cofax.org
Documentum	Web publisher	http://www.documentum.fr
Fatwire	Content Server	http://www.fatwire.com/fr/index.html
Filenet	P8 Web content manager	http://www.filenet.fr
IBM	DB2 Content Manager	http://www.ibm.fr
Instranet	Content Application Foundation	http://www.instranet.fr
Interwoven	Interwoven 6 Product Suite	http://www.interwoven.com
Jahia	Jahia 4.0	http://www.jahia.org
Microsoft	Content management server 2002	http://www.microsoft.com/france/cmserver/
OpenCMS	OpenCMS 5.0.1	http://www.opencms.org
RedHat	Enterprise CMS	http://www.redhat.fr
Spip	Spip 1.7	http://www.spip.net/fr
Tridion	Tridion R5	http://www.tridion.com/fr
Typo3	Typo3 3.6.0	http://typo3.org/
Vignette	V7 Content Services	http://www.vignette.com/fr
Zope	Zope 2.7.0	http://www.zope.org

(1) Nous indiquons prioritairement le nom du module dédié à la gestion de contenu web.

◆ Annexe 2 : Sources

ADAE. – Référentiel accessibilité des services Internet de l'administration française
www.adae.gouv.fr/article.php3?id_article=246

Patrice Bertrand. – Content management : les solutions open source (white paper). – Paris : Smile, 2003.

Antoine Crochet-Damais (coord.). – Numériser, gérer et publier ses contenus [dossier]. – *JDNet Solutions*, août 2004.
http://solutions.journaldunet.com/dossiers/webcontent/sommaire.shtml

Antoine Crochet-Damais. – Systèmes de gestion de contenu : la consolidation du marché devrait se poursuivre en 2004. – *JDNet Solutions*, 12 janvier 2004.
http://solutions.journaldunet.com/0401/040112_cms.shtml

Russel Nakano. – Gestion de contenu web : une approche collaborative. – Paris : Vuibert, 2002. – 229 p. – (Entreprendre Informatique)

Mark Pilgrim, Karl Dubost (trad.). – Plongez dans l'accessibilité : 30 jours pour rendre un site web plus accessible
www.la-grange.net/accessibilite

Michel Remize. – Geide et Content management : le contenu n'a plus de frontière. – *Archimag*, mars 2002, n° 152.

Olivier Roberget. – La gestion de contenu d'entreprise attend toujours son logiciel miracle. – *01 Informatique*, 11 octobre 2002, n° 1695.

W3C, J.J. Solari (trad.). – Recommandation CSS2 du W3C en version française
www.yoyodesign.org/doc/w3c/css2/cover.html

W3C, Service d'information du gouvernement (trad.). – Directives pour l'accessibilité aux contenus Web (version 1.0)
www.la-grange.net/w3c/wcag1/wai-pageauth.html