Le document numérique dynamique : une « étoile filante » dans l'espace documentaire

Katarzyna WEGRZYN-WOLSKA
Esigetel, école supérieure d'ingénieurs en informatique et génie des télécommunications

Résumé :

La majorité de pages Web existantes actuellement sont des pages Web créées dynamiquement. Ces documents qui n'existent pas réellement, sont créés pour une demande individuelle (automatique ou manuelle) et ils disparaissent après leur consultation. Cet article s'intéresse aux problèmes de la durée d'existence, d'accessibilité et d'archivage de ces pages. Les différentes définitions, catégorisation des documents dynamiques et leur mise en œuvre sont introduites dans un premier temps pour ensuite analyser les résultats de différents tests statistiques effectués dans l'objectif d'évaluation de durée de vie de documents numériques dynamiques.

La taille de l'espace documentaire exploitable dans la forme numérique augmente de plus en plus rapidement. Avec cette croissance nous observons une diversification de formes et formats des documents numériques : les documents statiques et dynamiques, les pages Web très variées et les documents multimédia sur les différents supports. Cet article aborde les problèmes de la durée de vie, de l'actualisation et d'archivage des pages Web dynamiques.

Caractère de documents dynamiques et ses différentes définitions

Document électronique et ses définitions

Avant de définir le terme « document dynamique », il est souhaitable de préciser la signification des termes *document* et *document électronique* comme tels. Voici la définition donnée par le **Centre ATO** (UQAM) et l'EBSI (Université de Montréal)[1] :

> **Document :** Désigne une entité identifiée et structurée contenant, entre autres, textes, tableaux, images et sons, pouvant être un objet d'étude, de traitement manuel ou électronique (par exemple, l'archivage), et d'échange entre des utilisateurs. C'est donc une entité constituée d'un contenant et d'un contenu ; ce dernier ayant surtout la caractéristique d'être communicatif au niveau social. Ainsi, par exemple, une pierre en

[1.] Centre ATO de l'université du Québec à Montréal (UQAM) et L'école de bibliothéconomie et des sciences de l'information (EBSI) de l'université de Montréal, Glossaire des termes d'ATO. <http://www.ling.uqam.ca/sato/glossaire/index.html> [février 2007].

> soi ne peut constituer un document ; par contre, une pierre gravée d'écriture peut en constituer un. De même, du texte, de l'image ou encore du son sur un support électronique peut constituer un document. Dans ce cas, on parle de *document électronique.*

Figure 1. Juste la pierre… ou document ?

Images extraites de : http://www.malexism.com/medias/rosette.html ; http://www.pinczow.com/muzeum/images/kamien.jpg ; photothèque de l'auteur.

Il est tout à fait intéressant de poser la question sur ce qu'est le document électronique dynamique. Est-ce que c'est un vrai document ou juste une présentation temporaire de données ? Est-ce que les documents dynamiques sont des documents créés de façon automatique, ou transformés en fonction des actions de l'utilisateur ? Le terme « dynamique » en tant que tel, est utilisé à plusieurs titres : pour les documents contenant différents moyens HTML dynamiques comme les « calques », *scripts*, etc., mais le terme de pages dynamiques correspond davantage à des pages construites « à la volée » sur le serveur. Quelle est donc la signification du terme « document dynamique » et la définition utilisée dans cet article ? Cet article analyse en particulier les documents créés en ligne sur le serveur.

Le terme « document dynamique » (synonyme « page dynamique ») est défini par la Banque de terminologie du Québec, dans *Vocabulaire d'Internet* accessible en ligne[2].

> **Document dynamique :** (page dynamique) c'est une page Web créée en réponse à la demande d'un utilisateur, dont la forme est fixe et le contenu variable, ce qui permet ainsi de l'adapter aux critères de recherche de celui-ci.

[2.] Banque de terminologie du Québec Vocabulaire d'Internet <http://www.olf.gouv.qc.ca/ressources/bibliotheque/dictionnaires/> [février 2007].

Mise en œuvre de document dynamique

Le document dynamique est créé en ligne. Un serveur Web (serveur HTTP) répond à une requête HTTP en renvoyant une page Web qui peut être une page statique ou dynamique. Lorsque la requête contient la demande d'une page dynamique (par exemple avec des données du formulaire en ligne), le serveur Web transmet toutes les données à une application (programme) demandée en vue de leur traitement et de la création de la réponse (page Web créée comme résultats d'exécution de programme demandé). Ensuite le serveur Web renvoie cette réponse sous la forme de page Web.

Différentes catégories des documents dynamiques

Existe-t-il différents types de documents dynamiques ? Pour répondre à cette question, il faudra prendre en compte aussi les différents aspects de leur création. Les documents dynamiques peuvent être construits sur la demande individuelle de l'utilisateur en fonction de ses requêtes (résultats de recherche sur le moteur de recherche, réponses à partir de données dans le formulaire, etc.) ou ils peuvent être créés ou modifiés automatiquement par l'application spécialisée (comme les différents sites d'actualité, forums de discussions, etc.). En conséquence, le comportement et la caractéristique de ces deux types de documents sont différents. C'est pourquoi les deux catégories de documents dans l'article sont traitées séparément. La première catégorie de documents créés sur la demande particulière de l'utilisateur est analysée et présentée sur l'exemple de pages de réponses venant de moteur de recherche. La deuxième catégorie, les documents créés automatiquement, est présentée sur l'exemple de pages venant de différents services de news et de sites de Weblogs.

Durée de vie et âge de documents dynamiques

Comment analyser la durée de vie de documents dynamiques si les documents dynamiques n'existent pas réellement, et s'ils disparaissent de la mémoire de l'ordinateur après leur consultation ? Dans l'article, la durée de vie de ces documents est considérée comme le temps pendant lequel les réponses à la même requête ne sont pas changées. En plus, c'est ce temps qui est visible pour l'utilisateur, puisque pour lui la différence en consultation de ces deux types de documents (dynamiques, statiques) est transparente. L'utilisateur dans son navigateur ne fait pas de distinction sur la manière dont le document consulté a été créé.

La deuxième question qui vient automatiquement est la question suivante : comment préciser l'âge de document dynamique ? Par l'exemple est-ce que la valeur indiquée dans l'en-tête http : *Modified* et *Expired*, ou dans *Meta Tag Expires* d'un fichier HTML, indique vraiment quand le contenu du document a été modifié, et quand ce document doit être considéré comme expiré ?

Dans cet article les problèmes de durée de vie, d'accessibilité et d'archivage de documents dynamiques sont présentés séparément à travers des exemples de sites (créant les pages dynamiques) d'actualité, de *Weblogs* et de moteurs de recherche.

Les sites d'actualité (*News*)

Sur le Web, il existe beaucoup de sites offrant le service d'actualité (*News*). Même s'ils publient toutes les différentes informations d'actualité et les dépêches de presse, les informations diffusées sur ces sites sont diversifiées (Christophe Asselin, 2004). Il existe les sites généraux comme l'actualité mondiale et l'actualité régionale, mais aussi les sites d'actualité dans des domaines plus précis. La majorité de tous ces services est créée automatiquement, les informations d'actualités sont mises à jour instantanément, sans interruption tout au long de la journée, ainsi le lecteur de ces services retrouve des nouvelles de dernières minutes chaque fois qu'il consulte la page de *news*. Par contre, il est souvent possible d'accéder aux anciens articles à partir d'archives disponibles sur leurs sites. La durée d'archivage pour différents sites est assez variée. Le tableau 1 présente les valeurs comparatives de temps de mise à jour et de la durée d'archivage pour les différents services d'actualité. Ces valeurs estimées auparavant par des tests effectués sur les sites concernés, ont été confirmées par les réponses reçues de la part de différents sites interrogés ensuite.

Les sites des Weblogs

Weblog (en anglais : *log, weblog* ou *blog* ; en français : « blogue » ou « joueb ») est un journal mis à jour régulièrement, sous la forme d'une page Web évolutive, présentant des informations de toutes sortes, généralement des pages dynamiques contenant des messages mis à jour régulièrement (Rebecca Blood, 2002 et Stephanie Booth, 2002**)**.

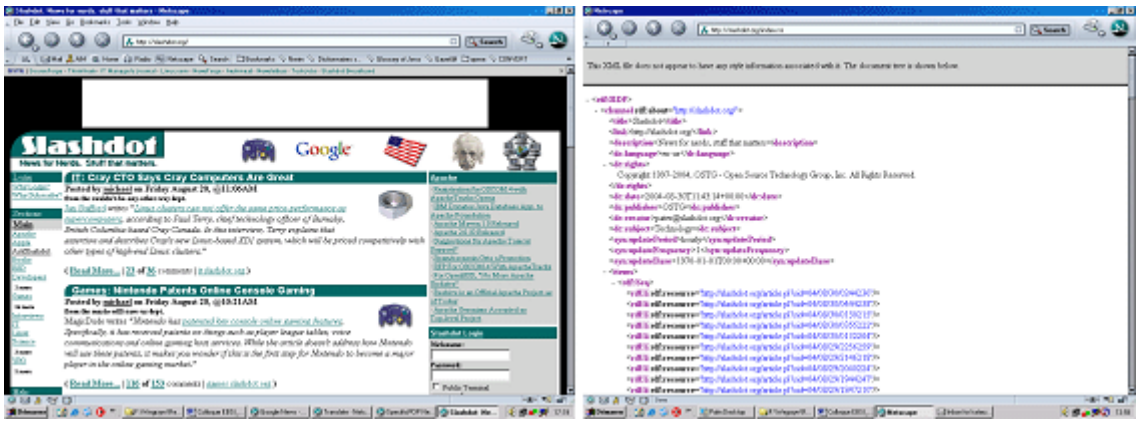

Figure 2. Exemple de page de weblog Slashdot (page HTML et son fil RSS).

Service de news	url	Actualisation	Archivage
Google français	<http://news.google.fr>	~20 min	30 jours
Google	<http://news.google.com>	~20 min	30 jours
Actualité Voila	<http://actu.voila.fr/>	1 jour	1 semaine
Dépêche Voila	<http://actu.voila.fr/Depeche/>	en temps réel (~30min)	1 semaine
CNN	<http://www.cnn.com/>		
Yahoo!News	<http://fr.news.yahoo.com/>	instantanément	1 semaine
TF1 news	<http://news.tf1.fr/news/>	instantanément	
News now	<http://www.newsnow.co.uk/>	5 minutes	
Les Infos	<http://www.lesinfos.com/>		À partir de 2000
CategoryNet Portail de la presse et des RP	<http://www.categorynet.com>	Chaque jour : lundi – jeudi pas actualisé : vendredi - dimanche (sauf info. brûlante)	indéfiniment
CompanynewsGroup l'information officielle des sociétés	<http://www.companynewsgroup.com>	en temps réel, en moyenne 40 communiqués par jour	2003 et 2004 archivés ; 1999 – 2003 en projet

Tableau 1. Paramètres comparatifs (rafraîchissement et archivage) pour les différents services d'actualité.

Moteurs de recherche

Les réponses fournies par les moteurs de recherche sont des pages de réponses dynamiques, créées en ligne. La durée d'existence (l'accessibilité) d'une page de réponses, c'est-à-dire le temps pendant lequel le moteur de recherche fournit la page identique, dépend bien sûr des réponses retrouvées par le moteur de recherche dans sa base d'index. Il est alors bien corrélé avec la fréquence de mise à jour de la base d'index. Le Tableau 2 contient les exemples de valeurs de temps de mise à jour d'index pour les différents moteurs de recherche.

Moteur de recherche	URL	Mise à jour d'index	
Google	<http://www.google.com> <http://www.google.fr>	4 semaines mais certaines pages sont rafraîchies quasi quotidiennement	
All the Web	<http://www.alltheweb.com>	très fréquente, depuis le printemps 2004 index commun avec Yahoo!	indique la date de visite par les robots
AltaVista	<http://fr.altavista.com/>	depuis le printemps 2004 index commun avec Yahoo!	

Tableau 2. Paramètres comparatifs (mise à jour des bases d'index) pour les différents moteurs de recherche.

Archivage

Analysant le problème de documents dynamiques il faut aussi poser la question : quelle est **l'accessibilité** et comment sont **archivés** les documents dynamiques, documents qui disparaissent après la consultation ? Les possibilités d'archivage sont très différentes. Les documents numériques dynamiques peuvent être imprimés (matérialisation de document, processus contraire de numérisation) ou sauvegardés par leurs demandeurs ou par les différents systèmes de caches et d'archives spécialisées. Il existe beaucoup de différents outils qui archivent l'image du Web actuelle (par exemple *Wayback Machine* de *The Internet Archive*[3]*)*. Ces outils essaient de retrouver et d'archiver toute la partie du Web visible (Steve Lawrence, 2001).

Figure 3. Exemple de pages d'archivage de *Wayback Machine* pour le site de Google news et BBC.

Bien sûr c'est une tâche très difficile. La taille du Web et le dynamisme de changement sont tellement grands qu'ils rendent l'archivage complet de l'image du Web pratiquement impossible. L'exemple de l'archive effectuée par le *Wayback Machine* sur les pages d'actualité de GoogleNews et BBC est présenté sur la figure 3. La comparaison des archives existantes sur ce site avec les données des statistiques effectuées (figure 6, Tableau 1) montre que l'état d'archive présenté sur le site de *Wayback Machine* est bien loin d'être complet.

Expériences et statistiques réalisées

L'article analyse la fréquence de la mise à jour de base d'index de moteurs de recherche et méta-moteurs de recherche et les résultats de différents tests statistiques effectués sur les différents sites d'actualité et de Weblogs. Le choix de différencier les documents dynamiques en séparant les

3. <http://www.archive.org/index.php> [février 2007].

documents d'actualité des pages de réponses de moteurs de recherche est justifié, puisque globalement le temps d'existence de contenu des pages de réponses venant de moteurs de recherche est beaucoup plus grand.

L'évaluation de fréquence de mise à jour de base d'index de moteur et méta-moteur de recherche est basée sur les expériences acquises pendant les travaux de réalisation d'un outil ayant pour but de retrouver les documents de l'administration française grâce à la méthode de méta-recherche (Wegrzyn-Wolska, Katarzyna, 2001 et 2004). Ces expériences ont été réalisées entre autres pour évaluer la pertinence des réponses et pour valider les liens vers les documents réponses. Les données et les résultats obtenus peuvent être utilisés pour l'évaluation de la durée de vie des documents dynamiques, puisque tous les documents réponses fournis par les moteurs de recherche et méta-moteurs de recherche interrogés sont toujours des documents dynamiques dans le sens de la définition introduite dans cet article. Une méthode assez simple de l'estimation de fréquence de mise à jour de bases d'index est l'analyse de fréquence de passage de robots d'indexation utilisés par le moteur de recherche. L'exemple des données de passage de robots récupérés à partir de fichier log. est présenté sur la figure 4.

Robots/Spiders visitors (Top 10) - Full list - Last visit

25 different robots*	Hits	Bandwidth	Last visit
MSNBot	11630+247	338.68 MB	30 Aug 2004 - 10:57
Inktomi Slurp	2251+569	11.63 MB	30 Aug 2004 - 09:01
Googlebot	1994+144	74.80 MB	30 Aug 2004 - 07:51
WISENutbot	633+5	18.58 MB	30 Aug 2004 - 04:58
Unknown robot (identified by hit on 'robots.txt')	0+373	0	30 Aug 2004 - 04:54
Voila	165+208	965.00 KB	27 Aug 2004 - 14:50
Alexa (IA Archiver)	231+45	9.09 MB	30 Aug 2004 - 05:26
GigaBot	171+56	6.06 MB	27 Aug 2004 - 23:27
Unknown robot (identified by 'crawl')	191+9	1.24 MB	29 Aug 2004 - 14:35
AskJeeves	120+19	6.28 MB	27 Aug 2004 - 18:50
Others	437+112	4.91 MB	

* Robots shown here gave hits or traffic "not viewed" by visitors, so they are not included in other charts. Numbers after + are successful hits on "robots.txt" files.

Figure 4. Fréquence des visites de robots d'indexation de moteurs de recherche.

Pour évaluer le temps d'existence de pages dynamiques, certains tests statistiques ont été effectués sur les pages de sites d'actualité (analyse de fil rss). Les statistiques effectuées et l'analyse des résultats obtenus montrent que le comportement de tous les sites testés est assez varié. En conséquence, les valeurs de temps d'existence de page sur tous ces sites sont également variées (figure 10, figure 11, tableau 3). Quatre catégories de sites différents ont été analysées : la rubrique JO de *Sportstrategies* (service d'actualité dans le domaine de sport), l'actualité française sur le site de TF1, l'actualité mondiale sur le site de BBC et le site de Weblog (Slashdot.org).

Sportstrategies est un exemple de site pour lequel les modifications d'actualité sont très régulières, avec un temps de rafraîchissement stable. Dans le cas de *Sportstrategies* les actualités sont diffusées régulièrement toutes les heures (figure 5).

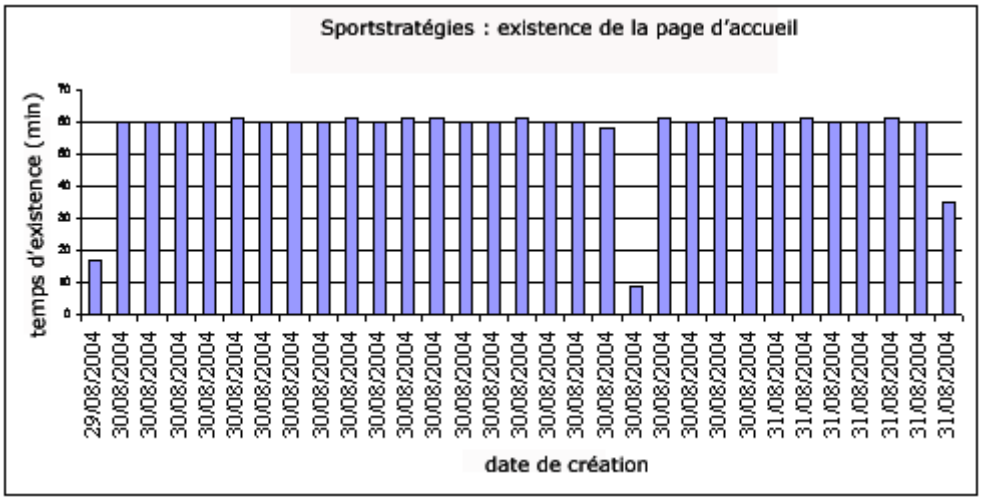

Figure 5. Durée d'existence de page d'actualité des JO 2004 à Athènes diffusée par le site *Sportstrategies*.

L'actualité mondiale sur le site de BBC est diffusée en temps réel. Les valeurs de temps d'existence des pages assez court et irrégulier (figure 6) prouvent le dynamisme des actualisations des articles diffusés instantanément.

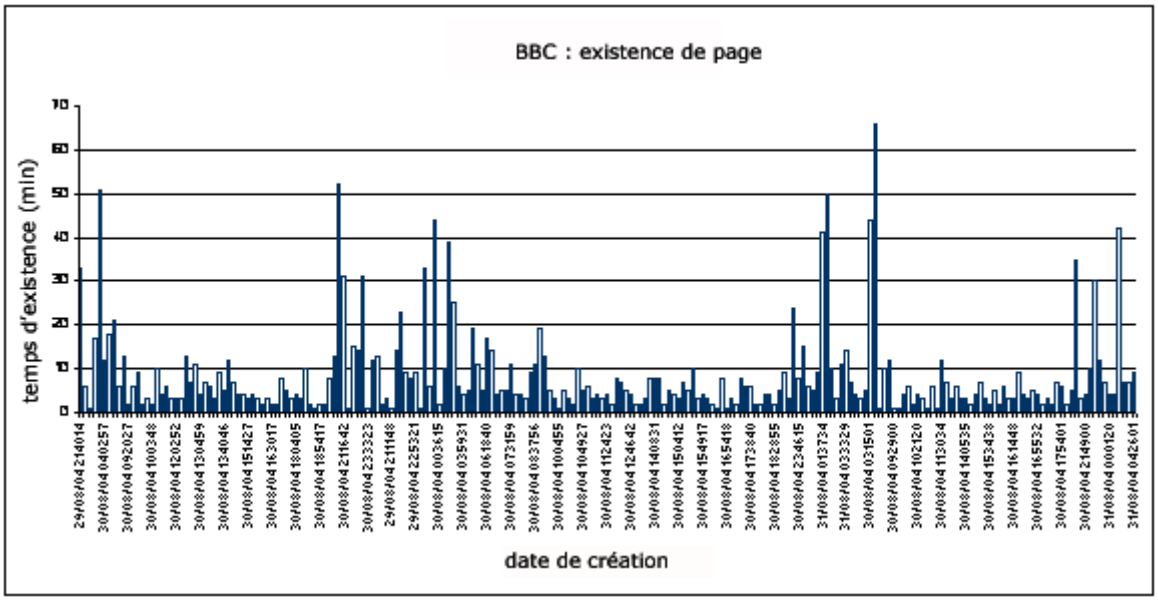

Figure 6. Durée d'existence de page d'actualité de service BBC.

Les articles et dépêches de presse diffusés par le site de TFI sont actualisés assez fréquemment pendant la journée. Par contre, il n'y a aucune modification de contenu de page d'actualité durant la nuit. Les temps d'existence de page pour ce site sont très différents dans ces deux cas. Pour bien illustrer les différents comportements du site TF1 pendant la journée et la nuit, ces deux situations sont présentées séparément (figure 7 et figure 8). Les deux « pics » à l'extrémité du graphe présenté sur la figure 7 correspondent aux pages d'actualité testées durant la nuit (page sans changements).

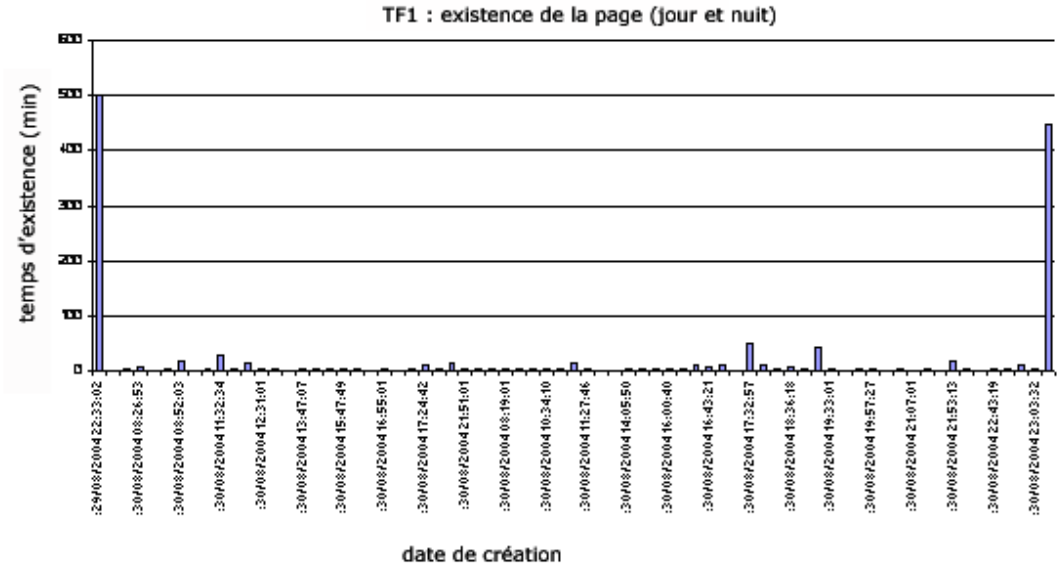

Figure 7. Durée d'existence de page d'actualité de service TF1 Actualité 24 heures/24.

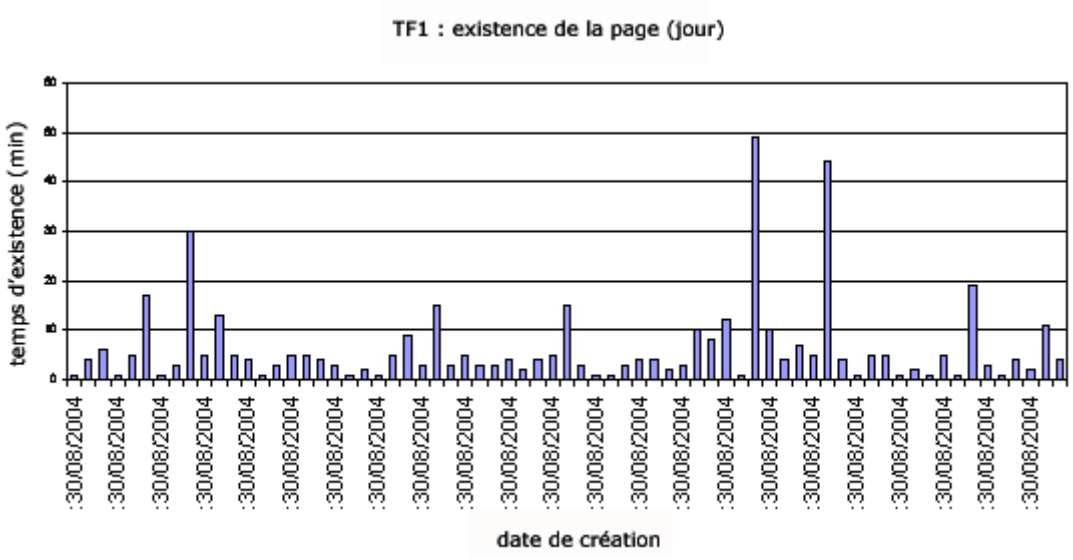

Figure 8. Durée d'existence de page d'actualité de service TF1 Actualité pendant la journée.

Le dernier site analysé est le site de *Slashdot.org*. Ce site de *weblogs* collectifs est un site de référence pour tous les fans de l'informatique et en particulier de logiciels *open source*. Les informations changent rapidement, les nouveaux articles sont proposés très souvent et la discussion sur les thèmes actuels est pratiquement sans arrêt, en continu pendant la journée… et la nuit aussi. Il n'est pas donc très étonnant que le temps d'existence de la même page sur le site *Slashdot* soit extrêmement court (figure 9), le temps moyen d'existence de la page est égal à 77 secondes (cf. tableau 3).

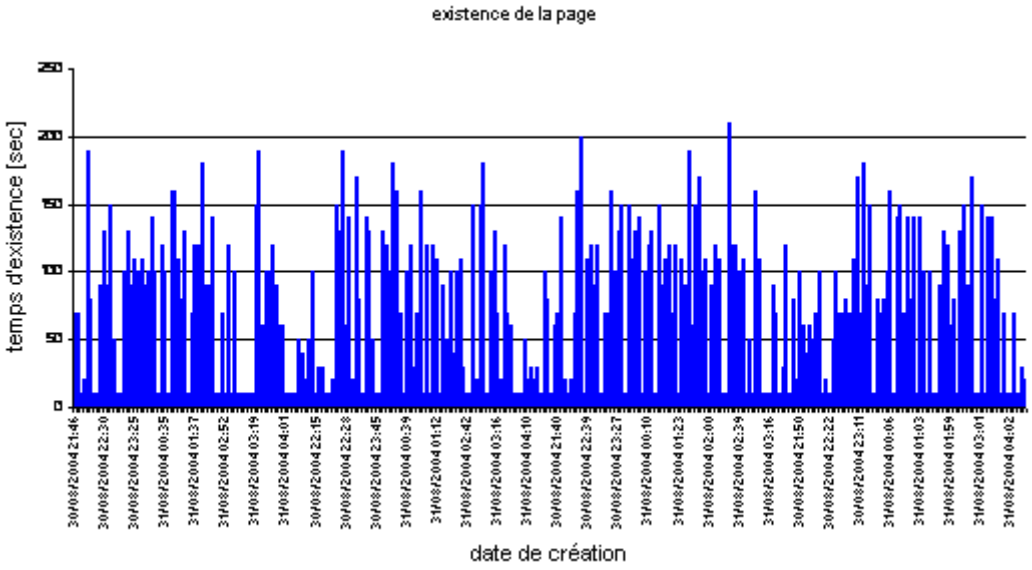

Figure 9. Durée d'existence de page de Weblog *Slashdot*.

Service testé	**Durée de vie**		
	Moyenne	*Min.*	*Max.*
Slashdot.org	77 secondes	10 secondes	22 minutes
BBC.News	8,5 min	une minute	66 minutes
TF1.actu (24/24)	19,5 min	une minute	502 minutes
TF1.actu (jour)	6,3 min	une minute	49 minutes
Sportsynergies	56 minutes	9 minutes	61 minutes

Tableau 3. Comparatif de temps d'existence de page pour les différents services testés.

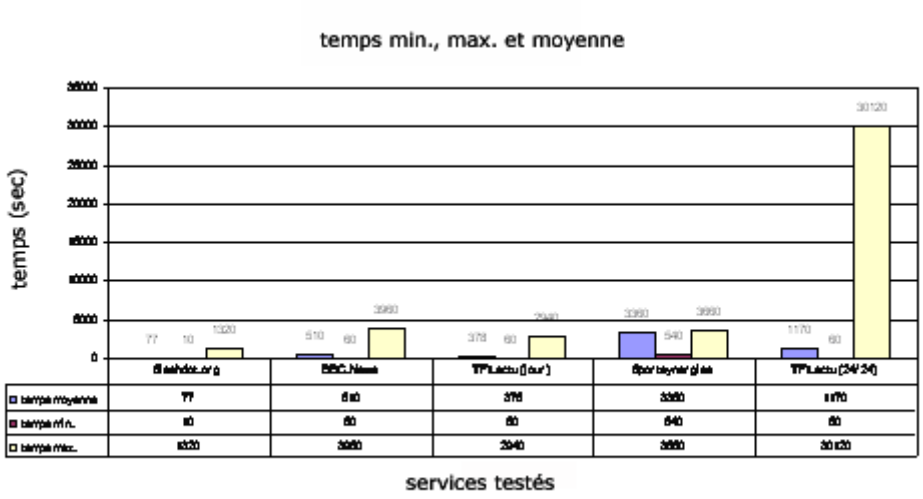

Figure 10. Temps min., max. et moyenne de l'existence de page pour les services testés (TF1 24/24).

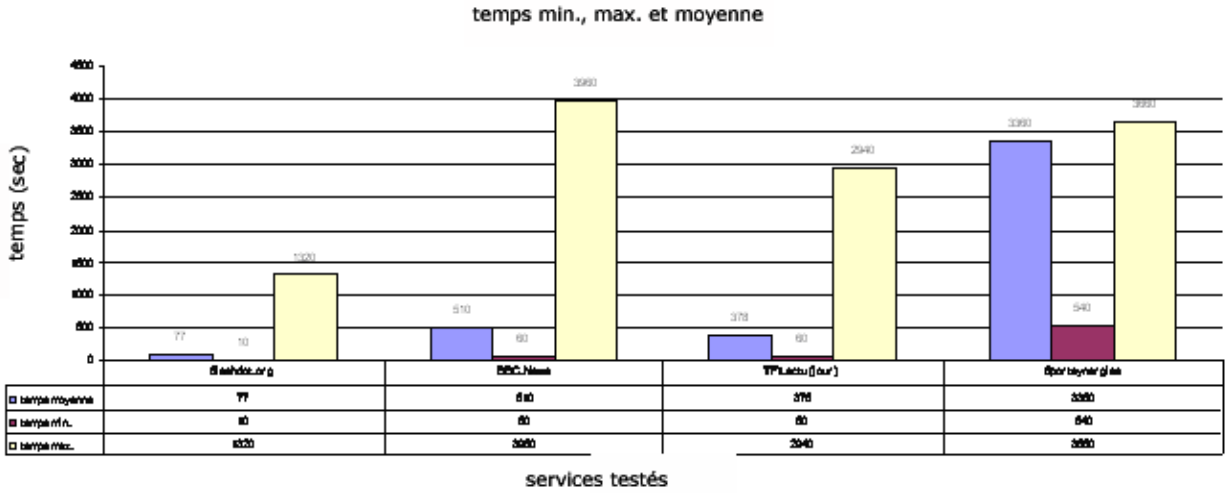

Figure 11. Temps min., max. et moyenne de l'existence de page pour les services testés.

Conclusion

Les documents numériques dynamiques n'existent pas vraiment, généralement ils disparaissent de la mémoire après consultation. Leur réelle durée de vie est donc très courte.

Par contre, les expériences effectuées montrent que les documents dynamiques restent accessibles beaucoup plus longtemps grâce aux différents systèmes d'archivage. La gestion de

durée de vie des documents dynamiques archivés devient identique à celle des documents statiques, puisque les documents dynamiques sont archivés sous forme statique.

Bibliographie

ASSELIN, C., *Chercher dans l'actualité récente ou les archives d'actualités françaises et internationales*, 2004, <http://c.asselin.free.fr/french/actua.htm> [février 2007].

BLOOD, R., *The Weblog Handbook: Practical Advice on Creating and Maintaining Your Blog*, Perseus Books Group, 2002.

BOOTH, S., *C'est Quoi Un Weblog*, 2002, <http://spirolattic.net/CestQuoiUnWeblog>, [2004].

LAWRENCE, S., « Online or Invisible? », *Nature*, Volume 411, Number 6837, 2001, p. 521. On-line version : <http://citeseer.ist.psu.edu/online-nature01> [février 2007].

WEGRZYN-WOLSKA, K., *Étude et réalisation d'un meta-indexeur pour la recherche sur le Web de documents produits par l'administration française*, Thèse de doctorat A/339/CRI, École des Mines de Paris, Décembre 2001.

WEGRZYN-WOLSKA, K., *FIM-MetaIndexer: a Meta-Search Engine Purpose-Built for the French Civil Service and the Statistical Classification of the Interrogated Search Engines*, WSS'04 The Second International Workshop on Web-based Support Systems avec le IEEE/WIC/ACM International Conference on Web Intelligence, Beijing, China, Septembre 2004.