Date submitted: 10/08/2010

Collecte des journaux en ligne : l'expérience d'archivage du *Sydney Morning Herald* (smh.com.au) à la Bibliothèque nationale d'Australie

Pam Gatenby
Directrice générale adjointe, Direction des collections
Bibliothèque nationale d'Australie
Canberra ACT, Australie
E-mail : pgatenby@nla.gov.au

[Cette présentation s'inspire largement d'une revue interne de projet préparée par Paul Koerbin, Responsable de l'archivage du web à la Bibliothèque nationale d'Australie. Qu'il soit ici remercié pour sa contribution.]

Traduction *:*
Philippe Cantie
(Bibliothèque nationale de France)

Meeting: **102. Newspapers**

WORLD LIBRARY AND INFORMATION CONGRESS: 76TH IFLA GENERAL CONFERENCE AND ASSEMBLY
10-15 August 2010, Gothenburg, Sweden
http://www.ifla.org/en/ifla76

Résumé :

La Bibliothèque nationale d'Australie a commencé à archiver une sélection de journaux en ligne au courant de 2009 pour les rendre accessibles par le biais de ses archives web (PANDORA). L'entreprise a soulevé de nombreuses questions, confirmant ainsi le fait que les journaux, notamment en ligne, constituent des entités complexes. Cet article présente et explique l'approche adoptée pour le Sydney Morning Herald (http://nla.gov.au/nla.arc-101523) et aborde les principaux problèmes rencontrés jusqu'ici : définitions et périmètre, particularités et caractéristiques des journaux en ligne, décisions sur la valeur de ce qui peut être collecté, contraintes techniques et informatiques, implications en termes de ressources. Cette présentation livre également les résultats d'une évaluation de notre approche actuelle qui doit avoir lieu mi-2010.

1 Contexte

La Bibliothèque nationale d'Australie a inauguré l'archivage sélectif de publications «numériques originales» en 1996 avec la création de PANDORA, les archives web d'Australie (http://pandora.nla.gov.au). Les critères de sélection de PANDORA donnent priorité aux ressources web présentant une importance nationale reflétant tous les aspects de

l'existence. Les publications en ligne dotées d'un équivalent papier ne sont pour la plupart pas collectées partant de l'hypothèse qu'ils ne font que dupliquer des contenus disponibles sous forme imprimée. Jusqu'à une date récente, ce critère de sélection était invoqué pour exclure les journaux en ligne mais il est aujourd'hui manifeste que la nature des journaux en ligne a connu des changements considérables.

Suite à une revue de nos activités d'archivage du web en 2008-2009, nous avons décidé qu'il était nécessaire de faire preuve de davantage de détermination dans nos efforts pour collecter des journaux en ligne. Jusque là, certains contenus de type journal avaient été collectés de manière ad hoc, dont des sites spécialisés d'informations et d'opinions se démarquant du courant dominant (http://pandora.nla.gov.au/tep/13027) et les blogs d'un certain nombre de commentateurs politiques.

2 Qu'est-ce qu'un journal en ligne ?

Un enjeu primordial de notre décision de collecter des journaux en ligne était de comprendre leurs caractéristiques et de savoir si le concept traditionnel de journal continue à s'appliquer dans le contexte en ligne. Les journaux en ligne traversent une période de changements particulièrement dynamiques et semblent poursuivre un certain nombre de visées. Voici quelques-uns de leurs traits communs :

- Ils comportent non seulement le reportage de nouvelles mais aussi un contenu en ligne substantiel sous la forme de billets d'humeur qui peuvent ne pas figurer dans la version imprimée, de blogs et de commentaires de contributeurs ;
- Ils fonctionnent comme des portails complexes vers une gamme de services et d'informations et utilisent les nouvelles technologies du web pour présenter le contenu sous une forme dynamique ;
- Leur contenu change régulièrement ;
- Ils peuvent enfin s'avérer très lourds en termes de taille de fichiers.

Le développement rapide des journaux en ligne laisse à penser qu'ils continueront à évoluer selon des modalités imprévisibles et qu'essayer par conséquent de les aborder de la même manière que les titres imprimés ne serait être d'une quelconque utilité. À des fins de collecte, il est peut-être plus pratique et opportun de cibler le contenu spécifique qu'ils publient comme par exemple les blogs, les commentaires et les « Unes » et tout autre aspect qui revêt une signification par rapport à la politique de collecte d'une institution. Dans la prise de décision, des considérations pratiques comme la fréquence à laquelle change le contenu d'un site doivent entrer en ligne de compte car elles risquent d'entraver une collecte efficace et appropriée, en fonction de l'infrastructure d'archivage numérique qui aura été mise en place.

3 Objet de la collecte

Au second semestre de 2009, la Bibliothèque nationale d'Australie a entamé un travail systématique de constitution d'une collection de contenus en ligne de type journal. Les journaux de communautés ethniques furent d'abord sélectionnés en appui à l'attention portée plus globalement à l'édition multiculturelle par la politique documentaire de la Bibliothèque. La collecte de la majorité d'entre eux ne présente aucune complexité. Ils sont d'une taille bien inférieure aux quotidiens, la négociation visant le permis d'archives est plus directe, et les contenus sont plus statiques. Au même moment fut obtenu auprès de la Fairfax Company le

droit d'archiver les contenus de la version en ligne du *Sydney Morning Herald* (smh.com.au), ce qui nous procura l'occasion d'aborder les questions de périmètre, de procédure et de technique de collecte que soulèvent de grands journaux en ligne caractérisés par leur nature véritablement complexe et dynamique.

Cette présentation porte essentiellement sur notre expérience de collecte du *Sydney Morning Herald* et sur les problèmes identifiés au terme d'une année d'archivage de ce titre.

Environ 45 titres figurent désormais dans la rubrique thématique « Journaux » de PANDORA (liste disponible à l'adresse http://pandora.nla.gov.au/subject/221). On compte dans le lot 22 journaux de communautés ethniques originellement numériques, 4 titres dont les journaux constituent la thématique principale (qui ne sont donc pas en soi des journaux) et un assortiment de publications de type journal centrées sur une thématique particulière.

4 Le Sydney Morning Herald (*smh.com.au*)

Le *Sydney Morning Herald* est l'un des titres de journaux les plus anciens et les plus réputés d'Australie. La Fairfax Company en est le propriétaire et l'éditeur depuis 1842. Le journal en ligne est un site complexe avec des articles et des blogs couvrant un grand nombre de sujets. Suite à des discussions avec le rédacteur en chef de la version en ligne, qui clarifièrent son mode de composition et les résolutions éditoriales qui en déterminent le contenu et vu la nature dynamique et complexe du site, la Bibliothèque décida de restreindre la collecte aux composants gérables et présentant à notre sens une valeur particulière d'un point de vue historique et culturel. L'importance accordée au journalisme d'opinion et aux blogs constitue un trait manifeste des journaux en ligne et nous avons décidé en l'occurrence qu'il devait être conservé. En outre, nous avons pris le parti de collecter et d'archiver la « Une[1] » de chaque numéro en ligne car celle-ci capte les nouvelles les plus importantes du jour et fait sens en elle-même. Les « pages » reflètent les décisions éditoriales concernant le choix des sujets et la présentation des contenus en vue de susciter l'intérêt des lecteurs en ligne et de capter leur attention. Il y a d'ordinaire quelque chose pour tout le monde. Cela peut aller de nouvelles sérieuses dans le domaine de la politique et de l'économie jusqu'à des articles grand public portant sur tous les aspects de l'existence.

Notre démarche de collecte

Le *Sydney Morning Herald* est collecté quotidiennement, sept jours sur sept, depuis juin 2009. Pour cet archivage journalier, un programme horaire a été défini dans PANDAS, le système de collecte et de gestion de contenu qui sous-tend PANDORA, de sorte que la capture des contenus intervient généralement très tôt entre minuit et deux heures du matin. Cette capture est celle qui est conservée les jours de week-end, mais les jours de semaine, une autre collecte est déclenchée manuellement à 10 heures du matin. Si cette dernière se déroule bien, c'est elle qui est conservée pour l'archivage et l'instance collectée aux premières heures du jour est supprimée de l'espace provisoire de stockage et ne fait donc pas l'objet d'un archivage. La

[1] Le concept de « Une » pour les titres de journaux en ligne ne va pas nécessairement de soi. Il s'agit manifestement de la page principale (ou du premier écran) mais est susceptible d'inclure également, à des fins d'archivage, le niveau de contenu situé à un clic de la page principale. Le titre et le premier paragraphe d'un article peuvent figurer sur la page principale mais la capture de la « Une » comprend la suite de l'article qui est en lien.

logique derrière cette démarche vise à capturer l'édition de milieu de matinée du journal tout en gardant la version disponible entre minuit et 2 heures du matin comme copie de secours, en cas de besoin.

La collecte est paramétrée pour capturer la « Une » du journal. Il ne s'agit pas comme nous l'avons vu d'un simple instantané du premier écran puisque la « Une » inclut également le contenu accessible en un clic. Le contenu en lien direct avec la première page est donc lui aussi collecté, ce qui signifie que, dans la plupart des cas, ce ne sont pas seulement les gros titres qui sont moissonnés mais bien l'intégralité des principaux articles.

La collecte quotidienne est de l'ordre de 4500 fichiers, pour un poids moyen de 50 MB et dure entre dix et 15 minutes. Bien qu'il soit difficile de dire ce qu'est un site de taille « moyenne » du point de vue de l'archivage, vu que ce qui est collecté dans PANDORA peut aller d'un simple fichier PDF à des sites très volumineux, on pourrait néanmoins considérer qu'il s'agit, pris isolément, d'un site relativement petit. Cependant, puisque sur une année, la collecte s'élève à presque 18 GB, on pourrait également considérer que son volume est très conséquent (étant donné que les fréquences de collecte les plus courantes sont annuelles ou semestrielles).

Pour ce qui est du temps de travail humain, le moissonnage de la « Une » ne requiert que 5 minutes vers 9h55 du matin pour lancer la requête manuelle dans PANDAS et le cas échéant, interrompre provisoirement les autres collectes en cours afin de permettre à celle du *Sydney Morning Herald* de démarrer, car PANDAS ne peut gérer concurremment que quatre collectes.

Le processus de contrôle de la qualité prend 10 minutes par jour. Il consiste à copier un certain nombre de fichiers préprogrammés dans les instances archivées pour résoudre les problèmes d'affichage ; puis à effectuer des contrôles aléatoires pour vérifier le résultat et l'absence de changements de format au sein du site. L'instance est archivée puis publiée sous une page d'entrée titre. La collecte de « Minuit » est effacée après la fin de celle de 10 heures du matin. Le lundi et le lendemain des jours fériés, il faut contrôler, corriger, archiver et publier les instances du week-end ou des jours fériés.

L'affichage du contenu archivé dans PANDORA se fonde sur une page de titre qui propose des informations comme par exemple le nombre d'instances collectées, l'identificateur univoque du contenu et sa disponibilité (ou non) dans le web vivant (pour un exemple de page de titre, voir http://pandora.nla.gov.au/tep/99834). En ce qui concerne le *Sydney Morning Herald*, un nouveau titre est créé dans PANDAS tous les mois sous la forme *Sydney Morning Herald* (juin 2009) ou *Sydney Morning Herald* (juillet 2009), afin de ne pas cumuler des centaines de liens dans de très longues pages d'entrée titre. Ces multiples titres sont classés dans la rubrique « Journaux » de PANDORA. Pour les regrouper et en faciliter l'accès, une « collection » a été créée dans PANDORA, avec une page d'entrée titre à la collection comportant les liens vers les pages d'entrée titre pour chaque titre mensuel (voir http://pandora.nla.gov.au/col/10281).

Les billets d'opinion, les analyses et commentaires associés aux journaux en ligne du Fairfax Network dont le *Sydney Morning Herald* et produits soit par des chroniqueurs du journal soit par des lecteurs, sont publiés en ligne en tant qu'entité séparée sous le titre *National times*. Cette collection est archivée tous les trimestres. Voir http://pandora.nla.gov.au/tep/100981

pour la page de titre dans PANDORA et le lien vers le détail des conditions d'utilisation fixées par l'éditeur.

Problèmes et contraintes

Les problèmes et contraintes exposés plus bas concernent notre expérience de collecte du SMH et sont par conséquent liés à la démarche choisie et à l'environnement systèmes au sein duquel nous opérons.

Les aspects techniques du processus d'archivage ne sont pas fondamentalement différents de n'importe quel autre titre collecté pour PANDORA. Comme la plupart des sites complexes collectés pour PANDORA, chaque instance nécessite bien un contrôle de qualité et un certain nombre de corrections manuelles. Les principaux problèmes relatifs à la collecte de la « Une » du SMH résultent de la fréquence, quotidienne, de la collecte.

Bien que déjà par le passé PANDAS ait été paramétré pour un programme quotidien de collecte, il s'agissait en général de collectes quotidiennes réalisées sur des périodes courtes et délimitées dans le temps, comme ce fut le cas lors de campagnes électorales. La technologie et la gestion de flux actuellement utilisées dans PANDAS présentent certainement des limites dans le cadre d'un processus continu et quotidien de collecte de journaux.

Horaire de programmation

Même si PANDAS comporte un programme préinstallé de collecte quotidienne, celui-ci ne peut être paramétré pour une heure déterminée. Comme toutes les collectes programmées, le programme quotidien débute aux premières heures du matin, ce qui constitue un bon horaire de collecte en raison de son faible impact machine mais pas nécessairement pour des contenus qui sont aussi liés à l'actualité qu'un journal en ligne. PANDAS permet certes plus d'une collecte par jour mais les collectes supplémentaires doivent être initiées en mode manuel. Actuellement donc, pour collecter une version de bonne qualité des contenus, il est nécessaire qu'un agent initie manuellement le processus en milieu de matinée, ce qui requiert à l'évidence une planification supplémentaire, en termes de personnel, pour s'assurer que chaque jour, un agent sera bien disponible pour effectuer cette tâche. Afin de réaliser les collectes du week-end et des jours fériés, et en guise de sauvegarde au cas où la collecte de milieu de matinée se solderait par un échec, il est nécessaire de maintenir le programme quotidien (de minuit). Cela signifie que cinq jours par semaine, une collecte de 50 MB est effectuée puis mise au rebut. Une possibilité serait de conserver cette collecte – même si l'on n'a pas à ce stade idée de la valeur qu'elle pourrait représenter.

Contrôle de la qualité

Les sites de journaux sont complexes et nécessitent des opérations correctives manuelles pour garantir la conservation de la mise en page. Vu l'objectif de collecter quotidiennement la vue de la « Une » du journal, la mise en page est considérée comme particulièrement importante. Nous avons fait preuve à cet égard d'une certaine efficacité en identifiant les éléments de style à corriger, en conservant copie de ces fichiers et en les appliquant à chaque collecte journalière. La difficulté est que l'agent doit garder l'œil sur d'éventuels changements de mise en forme et sur les modifications de ces fichiers « préprogrammés » qui doivent en résulter. Une telle situation s'est produite à deux reprises au moins depuis les débuts de la collecte du SMH.

Affichage

Comme il a été dit plus haut, l'affichage dans PANDORA ne se prête pas à l'affichage d'un titre collecté quotidiennement sur une longue durée. Cela aboutirait à une très longue page, avec en une seule année de collecte, une liste verticale de 365 liens. Pour régler ce problème, il fut décidé de créer un nouveau titre par mois ce qui implique la création chaque mois d'une nouvelle page d'entrée à ce titre. La page d'entrée titre conserve donc des proportions fonctionnelles. Cela signifie qu'il existe de nombreuses pages d'entrée titre mensuelles pour un titre donné, ce qui aboutit à une liste non chronologique lorsqu'on interroge l'index « Journaux. » Le problème d'affichage et la décision de créer un nouveau titre chaque mois engendre de toute évidence un surcroît de travail puisqu'il faut créer un nouveau titre dans PANDAS, placer des liens dans les pages d'entrée titre pointant vers les titres précédents et suivants et copier les fichiers préprogrammés dans le répertoire de la nouvelle instance.

Conclusion

Bien que l'archivage quotidien du *Sydney Morning Herald* soit géré au sein du système PANDORA tel qu'il existe aujourd'hui, l'infrastructure n'est pas particulièrement bien adaptée aux exigences d'un archivage quotidien du point de vue de la programmation horaire et de la gestion de la présentation. La complexité du site cible et la décision de capturer les contenus à une heure définie requièrent également une bonne dose de vigilance de la part des agents qui doivent être réactifs face à d'éventuels changements du site afin de maintenir la cohérence et les objectifs de l'archivage. L'expérience s'est avérée aisément gérable pour un grand journal en ligne au sein de l'infrastructure actuelle et le resterait pour un ou deux titres supplémentaires, mais il s'agit encore d'une démarche à petite échelle. Pour accroître le nombre de journaux à collecter, il serait nécessaire de s'attaquer aux limites actuelles de l'infrastructure existante.