



Programme National Américain de Numérisation de Journaux (PNNJ): une initiative à l'échelle du pays pour améliorer l'accès aux journaux historiques américains

Mark Sweeney
Librairie du Congrès
101 Independence Ave., SE
Washington, DC 20540-4760
mswe@loc.gov

Traduit par :
Simon Cadieux
Translation student
University of Montreal
Canada

Meeting: 99. ICADS

WORLD LIBRARY AND INFORMATION CONGRESS: 75TH IFLA GENERAL CONFERENCE AND COUNCIL
23-27 August 2009, Milan, Italy
<http://www.ifla.org/annual-conference/ifla75/index.htm>

RÉSUMÉ:

Cet article décrit les relations à partenaires multiples, les spécifications et outils techniques, la création d'accès et les éléments de préservation, qui sont inclus dans la conception du Programme National de Numérisation de Journaux aux États-Unis (PNNJ), un partenariat entre le Fond National pour les Humanités (FNH) et la Librairie du Congrès (LC). Le PNNJ est une initiative à long-terme dont l'objectif est de fournir un accès permanent à une collection nationale (en format numérique) de journaux à caractère bibliographique et de journaux historiques triés sur le volet, digitalisés par des récipiendaires du FNH dans tous les états et territoires américains. Ce programme offre à la Librairie du Congrès une zone d'essai pour le développement de programmes numériques à grande diffusion, et pour l'évaluation de besoins à long terme en gestion et en préservation de capitaux numériques. La phase de développement actuelle se concentre sur la création de pages de journaux numériques de substitution qui, par l'entremise d'un effort concerté, absorbent les objets numérisés en résultant dans un système, fournissent un accès aux données se voulant simple d'utilisation pour l'utilisateur, et implantent un système pouvant garder viable l'ensemble du contenu, pour un usage futur.

Le Programme National de Numérisation de Journaux aux États-Unis (PNNJ), un partenariat entre le Fond National pour les Humanités (FNH) et la Librairie du Congrès (LC), est une initiative à long-terme dont le but est de fournir un accès permanent à une collection nationale (en format numérique) de journaux à caractère bibliographique et de journaux historiques triés sur le volet, digitalisés par des institutions financées par le (récipiendaires du) FNH, dans tous les états et territoires américains. Ce programme capitalise sur l'exemple stratégiquement triomphal du Programme de Journaux aux États-Unis (PJEU, 1982-2009), financé par le FNH et avec le soutien technique de la LC – un excellent exemple d'une collaboration fructueuse, à la fois sur le plan national et étatique, dans le but de faire l'inventaire, cataloguer, et préserver sur microfiches un corpus soigneusement choisi de documentation à risque, à base de journaux. Le nouveau PNNJ ne fait pas seulement qu'accroître l'utilité des avoirs bibliographiques et microfilmés du PJEU, en améliorant l'accès à cette précieuse information, mais donne par ailleurs la possibilité à de nombreuses institutions de contribuer, par du contenu de journaux soigneusement choisi et digitalisé, à une ressource journalistique nationale et libre d'accès pour tous.

Les journaux historiques sont la source principale de comptes rendus d'événements ayant contribué au développement des communautés. Ils représentent un moyen par lequel les faits et les opinions sur des moments dans le temps, des personnalités d'importance, et des perspectives locales, peuvent être partagés — une ressource unique pour l'enregistrement et la compréhension des effets produits à la fois par des voix individuelles et collectives sur les idées, les événements, et sur l'identité démocratique, ainsi que pour la définition des archives historiques. Au cours des dernières décennies, sous le PJEU, la préservation de journaux sur microfilm et la mise en place de normes pour l'imagerie et la bibliographie furent des éléments importants des programmes d'archivage ayant à gérer et sauvegarder la vaste quantité de matériel nécessaire à une représentation efficace des archives historiques. Cependant, même cet aspect critique de l'éducation du bibliothécaire de journaux n'offre que peu de réponses aux questions soulevées par les besoins d'utilisation et d'accès à du matériel journalistique intensivement textuel. L'utilisation de cette précieuse ressource, archivée sur pellicule ou sur le papier original, est un défi de taille à la fois pour les bibliothécaires et pour les usagers, si l'on considère ses aspects physiques encombrants, du papier décoloré et cassant, ainsi qu'une organisation complexe. Même avec les meilleures normes d'archivage et les processus les plus efficaces, le contenu intellectuel du journal est réparti sur une présentation compliquée, avec des indices visuels variables et des petits caractères, épuisants pour l'oeil comme pour le mental. Le développement de nouvelles technologies de numérisation, de reconnaissance de textes, d'engins de recherche, etc., permettent maintenant au PNNJ de fournir un accès et une recherche améliorés pour ce matériel, ainsi que la supervision nécessaire à l'échelle nationale, afin d'établir les meilleures pratiques et normes pour la numérisation et la structure du matériel journalistique historique destiné à une ressource électronique de longue durée.

Puisque la collection nationale de journaux des É.-U. est dispersée dans des centaines de bibliothèques réparties dans tout le pays, un modèle de sélections décentralisées et de conversion numérique a été adopté, avec un agrégat de données fourni par la Librairie du Congrès pour l'accès et la préservation. Les objectifs principaux du programme sont à long terme – fournir un accès amélioré à des journaux triés sur le volet en créant et en regroupant des millions de journaux numérisés (et géographiquement diversifiés), et donner en même temps un nouveau mandat à des données bibliographiques et des avoirs déjà existants, comprenant plus de 139,000 titres, à l'intérieur d'un système permettant la recherche et le libre d'accès à tous.

Relations à Partenaires Multiples (Récipiendaires du FNH/LC):

Le FNH et la LC ont collaboré depuis 2004 au développement d'un programme, à l'échelle du pays, qui permet d'améliorer l'accès à ce matériel grâce à l'utilisation de nouvelles technologies et de canaux d'informations adaptés pour inclure un contenu représentatif de l'ensemble des états et territoires américains, contenu produit au cours de plusieurs décennies, et qui encourage l'interopérabilité entre les librairies numériques, par le biais du partage des spécifications. Un protocole d'entente entre le FNH et la LC définit très clairement les responsabilités des deux organismes, en ce qui a trait au développement général du programme national. Alors que le FNH gère et finance des compétitions annuelles entre organismes d'état, dans le but de sélectionner et de convertir des journaux historiques en format numérique, la LC concentre ses efforts au niveau des spécifications techniques du programme, de la gestion des données, et de la mise en service publique du contenu. Les organismes d'état, connus à l'intérieur du programme sous le nom de « récipiendaires », sont responsables de la sélection des journaux publiés à l'intérieur de leurs frontières, selon les critères du programme, ainsi que de leur conversion en un format numérique valide qui rejoindra un agrégat central à la LC.

En 2005, le FNH a réparti 1.9 millions de dollars entre six organismes – l'Université de Californie-Riverside, l'Université de Floride, l'Université du Kentucky, la Bibliothèque Publique de New York, l'Université de l'Utah, et la Bibliothèque de Virginie – afin qu'ils sélectionnent et convertissent un patrimoine de journaux représentant leurs collections d'état. Ces premiers récipiendaires furent choisis pour leur expérience des journaux historiques, de la numérisation de collections, et des infrastructures de librairies numériques. Lors de la phase initiale, le programme a produit un système de développement pouvant emmagasiner des centaines de milliers de pages de journaux historiques, convertis à partir des collections à la fois des récipiendaires de la LC et du FNH. En mars 2007, le PNNJ a lancé son service de diffusion par Internet à partir de ce système – *Consigner l'Histoire de l'Amérique* à <http://www.loc.gov/chroniclingamerica/>. Ce site fut lancé avec un répertoire-titre complet de journaux, des données créées sous l'égide du PJEU – 138,000 titres et 900,000 archives du patrimoine – et plus de 225,000 pages de journaux convertis, entièrement textuelles, recherchables, et publiées entre 1900 et 1910.

Le FNH a tenu une suite de compétitions dans le but d'augmenter graduellement la portée – à la fois géographique et temporelle – de la collection nationale agrégée, et afin de bâtir, au niveau de l'état, des compétences à grande échelle en numérisation de journaux. En 2007, le FNH a remis des prix à cinq des six organismes récipiendaires originaux, ainsi qu'à trois autres organismes – la Société Historique du Minnesota, l'Université du Nebraska, et l'Université du Nord du Texas. Cette tournée de prix allait se convertir en 800,000 pages de journaux, 100,000 pages par état, publiées entre 1880 et 1910. En 2008, le FNH a remis des prix à six organismes supplémentaires – le Département des Librairies de l'Arizona, les Archives et Archives Publiques, la Société Historique de l'Ohio, l'Université d'État de la Pennsylvanie, la Société Historique d'État du Missouri, l'Université d'Hawaï, Manoa, ainsi que la Librairie de l'État de Washington. Ces six organismes sont en train de convertir 600,000 pages de journaux, publiées entre 1880 et 1922, en provenance de leurs états respectifs. Une compétition en 2009, permettant d'étendre encore davantage le programme vers d'autres états et d'inclure une couverture chronologique plus grande (1860-1922), est présentement en voie de réalisation.

Spécifications Techniques et Outils:

Dans le cadre du développement et de la gestion générale du programme, la Librairie du Congrès fournit le support technique relativement au but premier du programme – la création d'un accès libre aux journaux historiques de la nation. Le rôle de la Librairie se divise en trois parties: établir les spécifications techniques de numérisation permettant l'agrégation, servir et unifier ce contenu par le biais d'un site Internet public, et sauvegarder de façon permanente le contenu agrégé. Alors que la LC passait en revue les moyens disponibles pour accomplir ces trois objectifs plus techniques, il devint évident que les pré-requis liés à l'objectif de sauvegarde du contenu allait influencer plusieurs des décisions prises au niveau des deux autres objectifs.

L'environnement du système du PNNJ, toujours en évolution, est basé sur la nécessité de soutenir quatre processus de travail majeurs, tels qu'identifiés dans le Modèle de Référence des Systèmes Ouverts d'Informations d'Archives (SOIA) : gestion de l'ingestion, de l'archivage, de la diffusion et de la préservation. Dès le début, la LC a reconnu la portée du programme tel que planifié – des millions de pages de journaux produites par plusieurs organismes différents, le tout échelonné sur une période d'environ 20 ans (l'équivalent, pour le moins, de centaines de teraoctets) – et l'engagement pris de gérer ces avoirs, par des organismes financés par le domaine public, a nécessité de mettre l'emphase sur la création d'avoirs numériques en accord avec les normes émergentes et les meilleures pratiques uniformisées. Des données bien formées, opérant dans une infrastructure technique robuste, représentent ce qui a été vu comme étant la meilleure approche pour assurer à long terme une gestion rentable du contenu.

Les premiers pas de la Librairie incluent l'élaboration de principes d'opérations de haut niveau, d'exigences de fonctionnement pour le système des avoirs numériques et du processus de travail pour la diffusion qui y est associé. Dans un climat d'émergence (et d'évolution) de pratiques optimales pour la préservation numérique, la LC a instauré une phase de développement explicite pour permettre la recherche et l'estimation des besoins en curation et en processus de travail à long terme, de même que des progrès incrémentaux vers la réalisation des objectifs du PNNJ. Les principes appliqués en effectuant des choix techniques avaient été prévus pour soutenir le développement d'un système viable, selon les meilleures estimations actuelles – libre, modulaire, certain de changer, et capable d'évoluer afin de pouvoir faire face à des utilisations futures.

De plus, les décisions prises l'on été en fonction de réalités liées à la structure générale du programme:

- Le contenu – versions analogiques (microfilm, papier) de journaux historiques – réside principalement dans des dépôts d'état, plutôt qu'à la bibliothèque nationale. Le programme requiert donc que les avoirs numériques soient distribués, une fois produits;
- Le financement nécessaire pour appliquer de nouvelles technologies permettant d'améliorer l'accès à ce matériel est limité, donc,
 - o considérant la quantité même de matériel disponible, le contenu à inclure dans le programme sera choisi, plutôt que de rendre disponible la totalité du corpus;
 - o les exigences techniques liées à la conversion du matériel devraient tenir en compte le potentiel, dans le temps, de réutilisation et de retraitement (scannez une fois, utilisez

- plusieurs fois)
- o devrait fournir un modèle pour des efforts de répartition semblables qui pourraient éventuellement interopérer – le partage des meilleures pratiques, les spécifications de conversion, et l’uniformisation de l’accès de base aux journaux historiques;
 - Démonstration d’un bon usage des fonds publics en permettant un accès libre et constant;
 - En prévision de changements à venir, ne refuser aucune option possible, ceci en développant un environnement de système qui serait libre, développable, et modulaire.

Dans le but de mettre sur pied une activité pouvant se développer et évoluer, le PNNJ a considéré de nombreuses exigences relativement à la production et la gestion de l’information numérique créée par les récipiendaires du FNH. Premièrement, afin de remplir son mandat d’agrégation et de gestion du contenu numérique à long terme, la LC se devait de respecter cinq exigences principales:

- adapter le contenu afin d’obtenir la meilleure qualité d’information, dans une optique de découverte et de réutilisation,
- assurer une technique uniformisée à l’ensemble d’un contenu créé au fil du temps par de nombreux producteurs,
- utiliser des formats ouverts et durables, pour encourager la préservation à long terme,
- développer une architecture des données qui permettrait à la fois une bonne gestion et une évolution dans le temps, et
- développer des processus de travaux évolutifs, des processus de gestion de la qualité pouvant supporter l’ingestion à grande échelle d’un contenu provenant de multiples producteurs.

Capitalisant sur son expérience de la numérisation à grande échelle des documents historiques, la LC a développé un ensemble de spécifications techniques basées sur les pratiques les plus efficaces, pour le contenu créé à même le PNNJ. Les spécifications d’image – TIFF, JPEG2000, et PDF – ont des assignations précises à l’intérieur du système du PNNJ (TIFF pour l’archivage, JPEG2000 pour la production, et PDF pour la manoeuvrabilité) et sont conformes aux pratiques courantes les plus efficaces, en ce qui a trait à la viabilité des formats de fichiers numériques. 1

Ces pratiques incluent une adoption par une vaste portion de la communauté de l’héritage culturel, la transparence de l’information numérique en soi, et l’auto-documentation à même le format de fichier. Les spécifications d’image pour le PNNJ, principalement une échelle de gris en 8 bit et entre 300 et 400 ppp, les efforts nécessaires pour capturer le plus de données possibles à partir des négatifs des microfilms de journaux, de manière à permettre un retraitement futur ainsi qu’une réutilisation à une date ultérieure en utilisant une technologie plus avancée. De plus, la LC a choisi un schéma standard de métainformations XML (Standard d’encodement et de transmission de Métainformations 2) pour la description des objets numériques au niveau de l’édition et de la pagination des journaux, et l’extension de schéma POTA (Présentation et Objet Textuel Analysés) 3 fut retenu pour la structuration des pages de texte lisibles à la machine et automatiquement reconnus (procédé connu sous le nom de reconnaissance optique de caractères (ou ROC). Les exigences des métainformations ont été conçues pour fournir un niveau d’accès de base aux pages de journaux, en prélevant

autant d'informations structurelles et techniques que possible, à partir du contenu des films comme du contenu intellectuel, en provenance du point de création numérique.

Le PNNJ a reconnu le fait qu'un modèle de production distribuée exigerait une amélioration des mécanismes pour assurer la qualité du contenu au moment de sa création et de son agrégation, de même qu'une incorporation explicite de métainformations destinées à fournir une aide à long terme pour la gestion et la viabilité des objets numériques. Ces exigences ont menées au développement de deux outils liés au PNNJ, une cible pour scanner de microfiches servant à l'analyse objective de la qualité d'image, et un logiciel de validation technique et de révision de qualité. Ces deux outils sont utilisés par des participants du programme dans le but d'assister dans la capture d'images techniquement valides et de haute qualité, et de garantir que les métainformations soient conformes aux spécifications techniques du PNNJ.

Les spécifications d'image du PNNJ visent à capturer la plus grande quantité de données possible, en provenance des microfiches de journaux, et le programme a établi ses spécifications techniques et les composants de ses processus de travail dans ce but. Quoique les éditions originales en papier puissent se substituer aux microfiches dans certains cas bien précis, il est entendu que les microfiches ont un rôle prédominant à jouer, car la plupart des éditions originales en papier, provenant de la période de temps désignée, ont souffert d'une détérioration significative ou ne sont tout simplement plus disponibles. La capture d'images provenant d'une cible standardisée, joint à un contenu visuel numérisé, est la pratique la plus efficace utilisée par de nombreux projets de bibliothèques numériques dans le but de faire progresser l'idéal que représente la production de matériel précis pouvant être géré en l'absence de l'article original. Reconnaisant le fait qu'une telle cible à l'essai n'existait pas à l'époque pour la numérisation de microfiches, le PNNJ a travaillé de concert avec les Associés de la Science de l'Image ⁴ afin de développer la Cible pour Scanner de Microfiche de Préservation (CSMP), une cible à l'essai standardisée sur microfilm (voyez la Figure. 1) qui, combinée au logiciel qui lui est associé, contribue à la création des images numériques de haute qualité que le programme requiert.

Figure 1. Cible pour Scanner de Microfiches de Préservation (CSMP), image fournie par les Associés de la Science de l'Image.

L'utilisation du CSMP par les participants au programme comble deux objectifs: créer un point de référence, et soutenir le contrôle continu de la qualité des images. Un groupe initial d'images cible scannées, provenant d'un dispositif de capture spécifique, peut être considéré comme un outil de référence pour anticiper le niveau d'efficacité de ce dispositif en particulier. L'analyse de ces images peut révéler si le système de capture, allant des optiques à la puce CCD et jusqu'au logiciel, est en mesure de créer des images rencontrant les spécifications du PNNJ. Si l'analyse de départ révèle un indice de performance du système inférieur aux prévisions, les résultats pourront amener l'opérateur à ajuster l'équipement pour générer un meilleur scan. La comparaison de scans de référence provenant d'équipements de balayage ou de vendeurs différents pourra contribuer à les départager.

Au cours de la production en masse de pages de journaux numériques, chaque titre, chaque pellicule et même chaque page possédera des caractéristiques visuelles différentes. Afin de développer un plan de contrôle de la qualité, l'utilisation de mesures objectives empruntées au CSMP peut se révéler très utile. Par exemple, si une image de page numérisée paraît floue, cela pourrait être une relique du processus de copie original, de l'état de l'original en format

papier à l'époque du tournage, du processus de conversion en microfiche ou de celui de numérisation. Déterminer quelle variable, parmi celles-ci et d'autres encore, est responsable des inquiétudes liées à la qualité peut être tout un défi. La capture et l'examen subséquent, à la fois visuel et par l'entremise du logiciel d'analyse, d'une image cible standardisée, capturée au même instant que la page de journal, peut fournir une mesure objective et non-rattachée-à-une-page-spécifique des facteurs de qualité de la numérisation au moment du scan. Si l'analyse de(s) image(s) cible indique que le scanner a exécuté sa tâche comme prévu, les divergences visuelles sont alors plus à même d'être repérées dans la microfiche, le scan étant une représentation précise de celle-ci.

La cible pour microfiche et le logiciel d'analyse qui lui est associé, développés pour le balayage de microfiches, contient une variété d'éléments permettant d'assurer le respect des spécifications d'imagerie les plus récentes émises par l'Organisme International de Standardisation (OIS). Ce qui suit est une description d'éléments relatifs à la qualité de l'image du CSMP, et comment ils peuvent être utilisés pour juger de la qualité d'un scan.

Généralement, la qualité d'image peut être divisée en catégories de reproduction tonale, de définition, et de reproduction du son et des couleurs. Puisque les microfiches sont conçues pour être monochromatiques, graphiquement parlant, la reproduction en couleur n'est pas une considération ici. Dans l'ISO 14524, la sensibilité aux tons du dispositif de capture se définit comme sa Fonction de Conversion Opto-Électronique (FCOE). La CSMP contient une série de cases grises, avec des degrés différents de noirceur, ce qui devrait être distinctement observable dans une image scannée de haute qualité. Ceci crée à la fois un indice visuel à l'effet que l'image cible soit parvenue à capturer le spectre complet de tons disponibles, et des points de données qui pourront être analysés par le logiciel, afin de calculer la FCOE pour le système.

Puisque les journaux contiennent un vaste assortiment de polices, formats de tailles, et éléments visuels en divers degrés de qualité, clarté et contraste, la reproduction des détails d'une image est essentielle pour capturer l'information contenue dans le journal d'origine. Il est non seulement important qu'une quantité suffisante de pixels par pouce soit capturée, mais également que le système optique puisse définir suffisamment de détails pour justifier ces pixels. Dans l'ISO 16067-2, la définition est mesurée dans le système par la Réponse de Fréquence Spatiale (RFS). La CSMP contient une bordure inclinée entre des zones blanches et noires, et des tableaux de résolution lisibles à l'oeil nu consistant en lignes étroitement espacées. Il s'agit là des données pour un logiciel servant à analyser le potentiel de RFS du système.

Le précédé de numérisation produit fréquemment des bruits imprévus: des reliques apparaissant au hasard ou systématiquement dans l'image numérique, et ne faisant pas partie de l'original. La CSMP mesure ces bruits en utilisant une série de carrés dotés de lignes verticales et horizontales très petites, ainsi qu'une longue ligne en diagonale sur toute la surface de la cible. Si de l'interférence est créée par la distance entre les pixels sur le détecteur et ces lignes, l'image cible la rendra visible en produisant un schéma de lignes largement espacées. La mesure de cette fluctuation est affichée par le logiciel d'analyse, qui fournit aussi de l'information sur l'amplitude probable d'une fluctuation acceptable.

Dans le cas des processus de travail liés à la production pour le PNNJ, l'analyse des images cible de la CSMP peut être accomplie aussi souvent ou aussi rarement que les gérants des projets récipiendaires le jugeront nécessaire. La cible peut être utilisée pour surveiller la

performance du balayage sur une base journalière, pour donner l'occasion de prélever des échantillons pour contrôle de qualité à partir d'une vaste quantité de données, ou elle peut être capturée et conservée pour analyse ultérieure, ou sur une base de nécessité. À ce jour, l'analyse de cibles scannées dans le contexte du PNNJ ont révélé de manière quantitative des performances de balayage dans les domaines de la reproduction tonale, de la définition et du bruit, améliorant la capacité des producteurs à surveiller les aspects du processus de numérisation. L'utilisation de la cible a contribué à la sélection de vendeurs. Elle a mis à jour les insuffisances de la performance de balayage. Tout aussi important: lorsque de piètres images de journaux, sur la microfiche originale, ont remis en question la validité du processus de balayage en soi, l'utilisation de la cible a également révélé à quels moments la performance de balayage fonctionnait correctement.

Afin d'assurer la conformité avec les autres spécifications techniques de métainformation décrites pour le PNNJ, un deuxième outil, utilisable par le personnel du PNNJ, les organismes récipiendaires, et les vendeurs de numérisation, est devenu nécessaire pour la validation des aspects techniques des données et des processus de garanties de qualité du PNNJ. Afin d'offrir un soutien efficace et évolutif, il fut évident que l'outil nécessaire aurait à soutenir un amalgame à la fois d'analyse automatisée de caractéristiques techniques, lorsque applicable, et de caractéristiques appropriées pour un contrôle humain supplémentaire. Une approche automatisée pourrait mesurer efficacement le degré de conformité technique (par exemple si un champ est composé de la bonne catégorie de données), alors qu'un lecteur d'objet de données supporterait mieux une inspection humaine subjective (par exemple si le champ de données est correct).

Pour soutenir une conformité technique automatisée, beaucoup de travail a déjà été accompli à l'université d'Harvard, par la création du logiciel JSTOR/Harvard d'environnement de Validation d'Objet à code ouvert (EVOJH) 5. Ce logiciel est en mesure de mesurer et de caractériser bon nombre de particularités des types de fichiers (JPEG 2000, PDF, TIFF) utilisés dans le cadre du PNNJ. De la programmation additionnelle fut réalisée à la LC pour développer les capacités du logiciel, et afin de valider la métainformation XML, essentielle au système du PNNJ 6. Ce code analytique fut joint à un interface graphique d'utilisateur qui devint connu sous le nom de Visionneur et Validateur Numérique (VVN). Le VVN a alors permis l'analyse automatisée du critère objectif des objets de données, garantissant que les bonnes catégories de données soient employées, ainsi qu'une révision de la qualité visuelle pour garantir que la métainformation soit utilisée correctement (par exemple, que le bon titre soit utilisé, que la date dans la métainformation concorde avec la date de l'image de la page).

Au cours de la validation, le VVN vérifie approximativement 100 caractéristiques du lot de données. Si les fichiers répondent aux spécifications, il extrait les données d'en-tête à partir des différents types de fichiers auto-documentateurs, pour fin de transformation en Stratégies d'Implantation de Préservation de Métainformation (SIPM) et de métainformation pour images en schéma XML schémas (MIX), à même l'objet METS associé. De plus, le VVN ajoute une signature numérique à l'objet METS pour chaque fichier associé. Cette signature numérique peut être vérifiée ultérieurement, afin de déterminer si le fichier a été modifié entretemps, soit volontairement, par erreur de l'opérateur, ou par dégradation des octets. Par conséquent, le VVN permet que la validité des fichiers de données du PNNJ soit contrôlée tout au long de leur cycle de vie.

Pour suppléer à ces outils déjà en usage par les récipiendaires et les vendeurs, le PNNJ a par ailleurs débuté l'utilisation de la panoplie d'outils et spécification BagIt 7, créée par la LC en

partenariat avec le Programme National de Préservation d'Infrastructure d'Informations Numériques (PNPIIN). Tôt dans le programme, le PNNJ a déterminé qu'afin de pouvoir maintenir un programme à la fois viable et rentable, les données produites par le PNNJ devaient être gérées de telle façon que leur fiabilité et leur qualité à l'usage soient garanties. Une des premières leçons fut d'identifier le besoin de minimiser l'interaction humaine (et donc l'erreur humaine) dans le cycle de vie des données. À cette fin, quand les données du PNNJ sont reçues par la LC, valides et avec une signature numérique pour chaque fichier intact, la livraison des données est alors "emballée" en utilisant la spécification BagIt afin d'automatiser la réception, le stockage et la récupération du contenu. Une fois emballé, des utilités de transfert général fourniront des services de gestion pour l'emballage, qui amélioreront la fiabilité, la trouvabilité, et l'intégration à d'autre matériel de collections numériques.

Les outils décrits ci-haut contribueront à assurer que l'investissement initial du PNNJ, en matière de spécifications de données et de contrôle de qualité, produise une ressource viable, dont la valeur surpassera les coûts de gestion et d'entretien à long terme.

Modèle de Système d'Accès:

Après un temps de recherche, la LC a développé des exigences de base pour un système d'accès, par un grand nombre d'exemples d'usage et de planification de scénarios. La fonctionnalité de l'accès de base pour le programme fut définie comme étant la capacité d'un utilisateur moyen à chercher et/ou feuilleter alphabétiquement les répertoires principaux d'archives de journaux, explorer divers titres numériques par date de publication ou par ordre logique de page, et à soutenir une recherche par mot-clé simple, au niveau de la page de journal. Le texte lisible à détection automatique (RCO), avec les données de repérage de mots associés dans le schéma ALTO, a fourni la structure de page de base et la capacité de recherche de mots requises, de même que de l'information professionnelle utilisée dans un interface visuel pour mettre en valeur les résultats de recherche. De la métainformation structurelle ou descriptive, identifiant des parties de page, ne fut pas incluse dans la spécification afin de maximiser les ressources disponibles et de répondre à la nécessité de fournir un accès de base au contenu. La vérification officielle de la capacité d'utilisation fut exécutée à l'aide d'un ancien prototype comprenant cette fonctionnalité, pour vérifier les hypothèses et besoins relatifs à un usage général.

Dans le développement initial du site, le site Internet d'accès en tant que tel (le logiciel de navigation) fut envisagé comme un élément constitutif du dépôt de la librairie, étroitement lié à la préservation et à la gestion des données du PNNJ. À ce moment-là, le cycle de vie préservé de l'avoir numérique était réalisé grâce à l'emploi d'outils de niveau système discrets et d'habiletés liées au système chez le personnel technique, pour les processus d'ingestion, d'indexation, et de diffusion à travers les services de dépôts liés au logiciel de navigation. Cependant, à mesure que le programme et l'infrastructure technique se développaient, la gestion technique et les stratégies d'accès de la LC évoluèrent. L'architecture de gestion des données fut ré-inventée afin d'exploiter plus efficacement à la fois les outils de publication pour sites Internet en pleine évolution, et les principes réalisables de gestion pour librairies numériques, ce qui eut comme résultat d'unir plus librement le site Internet aux systèmes liés à l'inventaire interne, le processus de travail, le stockage, et le transport.

En mai 2009 et à cette fin, le PNNJ a présenté une architecture de système et un logiciel d'accès révisés, transparents pour la plupart des utilisateurs individuels, afin de soutenir une variété de nouvelles fonctionnalités et de besoins en ressources. Les objectifs premiers de la révision furent:

- Améliorer la performance d'accès de manière à permettre un accès, par robots et moteurs de recherche, au contenu de *Consigner l'Histoire des États-Unis*. Fournir cet accès augmentera grandement le taux d'utilisation de la collection de journaux, en la mettant en face de millions de gens pour qui elle était auparavant voilée.
- Ajouter des interfaces standard de programmation (Logiciel d'Interfaces de Programmation, ou LIP) avec comme objectif d'améliorer la portée du site, quoiqu'actuellement plus pour les applications composites et les amateurs que pour les moteurs et robots de recherche. Par exemple, l'utilisation du LIP OpenSearch, qui permet aux utilisateurs d'explorer les pages et les titres de journaux directement à partir de leur navigateur de recherche, et de s'abonner aux résultats en tant que transmission. De plus, l'accès par LIP permet également d'utiliser et de ré-arranger le contenu en tant que tel, tout en prenant avantage du travail de modélisation et d'indexation déjà accompli dans la librairie.
- Réduire la complexité du code, grâce à l'étroite relation entre les services de préservation et d'administration, et avec un oeil vers des rénovations futures de l'apparence et de la maniabilité de l'interface d'accès. Cela donne un site ayant presque vingt fois moins de lignes de code que l'implantation originale.
- Dénouer les relations entre les composants d'accès et de préservation dans l'environnement de gestion des données, de façon à augmenter les services offerts par les deux, spécifiquement au niveau de l'augmentation des capacités de traitement des données du PNNJ. Ceci afin de rendre le processus d'acquisition, de gestion, et de fourniture d'accès aux données consistant, reproductible, vérifiable, et automatisé.

Le plus important: cette révision a été complétée avec succès, augmentant l'accès et l'utilisation du site par de multiples facteurs, sans aucuns changements aux spécifications techniques ou structures de l'objet numérique du PNNJ.

Les projets de programmes en code ouvert retenus pour la révision du logiciel d'accès incluent : le Serveur Internet Apache, la structure pour publication Internet Django 8 (provenant du secteur de publication de journaux et capable de combler plusieurs de nos cas d'usage en ressources numériques avec très peu d'efforts supplémentaires), la Librairie JQuery JavaScript, la base de données MySQL, le serveur de recherche Solr, ainsi que la librairie RDFLib Python. Ces outils fonctionnant en code ouvert deviennent rapidement très populaires au sein de la communauté des librairies numériques, à cause de leur flexibilité, fiabilité, et aise d'utilisation à la LC, à l'externe comme à l'interne. La version actuelle de l'interface d'utilisateur pour les données du PNNJ, baptisé : *Consigner l'Histoire des États-Unis*, (<http://chroniclingamerica.loc.gov>), est disponible en libre accès au public, sur le site Internet des Collections Numériques de la Librairie du Congrès.

Le site inclut actuellement plus de 1 million de pages de journaux, extraites de plus de 100 titres publiés entre 1880 et 1922, et provenant de 11 états et du District de Columbie. Un répertoire de journaux publiés aux États-Unis entre 1690 et aujourd'hui, de même que de

l'information sur les librairies qui les conservent sous forme physique et numérique, y est également disponible. Depuis mars 2007, le site a mis ce contenu à la disposition d'environ 400,000 visiteurs, avec plus de 6.8 million de visionnement de pages. En plus d'une fonctionnalité de recherche de base par mots-clé, le site Internet donne accès à de l'information de citation pour chaque page de journal, des calendriers visuels indiquant les éditions disponibles pour une année donnée, des fichiers pour téléchargement et réutilisation, des caractéristiques spéciales d'imprimerie afin d'imprimer des images détaillées d'une page, des liens « marque-pages » persistants pour chaque visite du site, un historique de journaux pour chaque titre numérisé, une diffusion hebdomadaire en RSS des points marquants du contenu des journaux et des développements du programme, et plus encore. Les récentes améliorations de la performance du site, décrites plus haut, encouragent l'utilisation du contenu, en plus de l'interface du site Internet de *Consigner l'Histoire des États-Unis*, en utilisant de nouvelles techniques de recherche telles que l'exploitation de données, la visualisation ou des méthodes de découvertes par utilisateurs de machines. Des domaines probables pour des améliorations futures au niveau de l'accès sont : l'incorporation en plusieurs langues du contenu de page et des procédés de recherche qui y sont associés, des analyses et manipulations automatisées supplémentaires du ROC afin d'améliorer la spécificité de recherche, de même que l'expérience générale comme utilisateur, ainsi que de la "capacité de partage" supplémentaire au niveau des données de page.

Rendre Viable le Contenu:

Un composant important de l'accomplissement du mandat de la LC dans ce programme est le développement d'un environnement de système pouvant garantir que les avoirs numériques, acquis de plusieurs sources différentes et sur une longue période de temps, seront viables. L'environnement doit être en mesure de garantir que lorsque les gens, les processus, et les technologies changeront, les avoirs numériques pourront (de manière transparente et automatique si possible) migrer d'une génération à l'autre. Une architecture de dépôt appropriée est un composant essentiel pour déterminer si un environnement de préservation numérique est conçu avec succès.

Le développement d'une infrastructure technique et de services à la librairie, dans le but de soutenir, fournir un accès, et améliorer la gestion à long terme des données du PNNJ, est un processus continu. Les deux couches architecturales majeures dans le dépôt sont : la couche de préservation, ou "archive" (stockage de données gérées), et la couche de données des gestion. La différence clé entre les deux couches se situe au niveau de la performance. La couche de préservation met l'accent sur la durabilité (ou longévité) des avoirs numériques préservés, alors que la couche de gestion des données met l'accent sur la vitesse d'entrée/sortie (E/S), la richesse de fonctionnalité, et la flexibilité de la gestion des données.

Tout au long de la courte histoire du programme, de réelles menaces à la préservation ont délimité les zones connues et, jusqu'ici, inconnues de risque dans l'acquisition, la gestion et la préservation fiables de ces données ². L'évaluation des leçons apprises, ainsi que l'apport provenant de d'autres projets de la LC, ont largement contribué au développement d'outils généralisés et de services supplémentaires qui amélioreront la fiabilité en ce qui a trait à la gestion des processus de travail complexes, associés à l'acquisition, la gestion, l'accès à, et l'archivage des données.

Amélioration récente faite à la trousse à outils du PNNJ, le logiciel de Transfert du PNNJ est une collection de trois outils: un interface de processus de travail, des mécanismes

automatisés de transport d'octets, et une base de données d'inventaire automatisée. Alors que les données traversent plusieurs étapes de leur cycle de vie, dans la librairie numérique, elles doivent être surveillées, contrôlées, récupérées et stockées, et nécessitent une connaissance en corrélation de ces processus, et plus encore. L'implantation initiale a uniquement permis au personnel technique de trouver, récupérer, et gérer les avoirs archivés en utilisant des outils de niveau système, privant le personnel curatorial et de gestion de programme d'un accès direct à la collection de librairie numérique. Afin de mitiger les risques associés à cette absence de connection, pour le PNNJ et d'autres projets, la LC a développé des outils de «transfert» permettant une administration plus robuste et rentable de l'information numérique, ainsi qu'une capacité d'évolution 10. Utilisés de concert, ces outils forment "l'arrière-plan" du site Internet, *Consigner l'Histoire des États-Unis*, d'une manière dissociée et flexible tout en rencontrant les critères de productivité générale et les objectifs de performance pour le programme, de même que sa croissance anticipée, à la fois en matière de contenu et d'utilisation.

Dans l'avenir, le PNNJ continuera à améliorer ces outils au-delà des processus de travail élémentaires, afin d'inclure une gérance de la conservation pouvant offrir des fonctions/aspects faciles à utiliser, pour créer, lire, mettre à jour, effacer, naviguer, surveiller, et faire des comptes-rendus du contenu de journaux, préservé de façon permanente en format numérique. Encore une fois, ces outils fonctionnels pourront sans doute être standardisés à d'autres activités de gestion de collections numériques, et seront d'une utilité sans cesse grandissante à long terme.

Soutenir l'Infrastructure pour la Viabilité:

Le FNH et la LC ont pris un engagement à long terme pour le développement du programme et de ses avoirs numériques, incluant une entente officielle concernant les objectifs du programme, le partage des coûts pour le développement et la gestion des produits du programme, et la direction en coopération du développement du programme. Afin de remplir son mandat de fournir un accès permanent vers ce contenu historique de grande valeur, la LC a mis en oeuvre le développement d'une infrastructure de soutien – à la fois programmatique et technique – dans le but de permettre la viabilité à long terme de la collection.

L'infrastructure établie à la LC inclut également une équipe de gestion du programme à l'interne, composée de parties prenantes représentant les intérêts des collections, la production numérique (conversion et acquisition), et la préservation numérique. Ces parties prenantes ont de l'expérience pratique dans un grand nombre de programmes à la LC, incluant le développement de collections de journaux, les collections historiques numériques de la Mémoire Américaine, des partenariats financés par Ameritech, la technologie de l'information et le Programme National de Préservation d'Infrastructure pour l'Information Numérique (PNPIIN). Ensemble, ces membres de comités représentant divers groupes de gestion au sein de la Librairie contrôlèrent avec succès les rôles et les documents à fournir de la LC, qui permirent de combler les attentes du programme – ils ont administré avec succès un modèle de production distribuée, fournissant un interface Internet pour acquérir des données, et développant un environnement de système pour maintenir et soutenir le contenu numérique.

Afin de pouvoir accomplir même un seul de ces objectifs, il fut essentiel que la LC mette également sur pied une équipe de développement technique zélée, représentant diverses spécialités – dont l'architecture de préservation et le développement de dépôts, le modelage

des données, le développement de logiciels, le développement de l'analyse de recherche et d'interfaces pour utilisateurs – et désireuse d'expérimenter et de contribuer à l'avancement des meilleures pratiques en matière de préservation numérique. Cette équipe a partagé son expertise (et parfois même son personnel) avec d'autres initiatives de la LC relativement aux dépôts, incluant les journaux électroniques, l'archivage Internet, ainsi que d'autres projets de collections numériques, en utilisant et en généralisant les leçons apprises lors du développement initial du PNNJ pour étendre leurs efforts, portant sur les dépôts, à d'autres types de contenu. Le groupe de développement technique qui soutient le PNNJ est impliqué non seulement dans la création d'un environnement de système ayant atteint les objectifs du PNNJ, mais également dans la mise sur pied d'un centre de développement de dépôts (matériel informatique, logiciels, et systèmes) à même la LC, dans un but de recherche continue sur les défis que représentent la préservation de tous les types d'informations numériques.

En conclusion, le PNNJ offre à ses participants l'occasion de travailler avec de multiples partenaires, ce qui a pour effet: la création d'un contenu de journaux numériques de choix, qui rencontre les strictes spécifications techniques pour l'agrégation, qui rencontre les besoins fondamentaux en accès d'utilisateurs au delà de ce qu'il est possible de faire en version analogique, et où les données sont emmagasinées dans un environnement de système ayant un haut degré de viabilité. Les décisions de front immédiates, concernant les pratiques et les stratégies les plus efficaces menant au succès du programme ont été validées. Alors que le programme continue de se développer et de prendre de l'expansion, la LC continue de s'adapter et de faire évoluer les outils et systèmes disponibles dans le cadre de ce programme. De faire face au défi de construire une collection nationale de journaux numérisés révélera à la face du monde les besoins et les possibilités pour la préservation de toute information numérique.

1 "Sustainability of Digital Formats – Planning for Library of Congress Collections".

<http://www.digitalpreservation.gov/formats/> , accès en date du 19 mai 2009.

2 Metadata Encoding and Transmission Standard (METS), <http://www.loc.gov/standards/mets/> , accès en date du 19 mai 2009.

3 Analyzed Layout and Text Object Schema (ALTO), <http://www.ccs-gmbh.com/alto/> , accès en date du 19 mai 2009.

4 Consultez : Image Science Associates, <http://www.imagescienceassociates.com/> , pour plus d'informations, accès en date du 19 mai 2009.

5 JSTOR/Harvard Object Validation Environment (JHOVE), <http://hul.harvard.edu/jhove/> , accès en date du 19 mai 2009.

6 Pour une explication plus détaillée sur les stratégies de validation d'objets numériques, implantées pour le PNNJ, consultez : Littman, Justin, "A Technical Approach and Distributed Model for Validation of Digital Objects", *D-Lib Magazine*, 12:5 (Mai 2006).

<http://www.dlib.org/dlib/may06/littman/05littman.html> , accès en date du 19 Mai 2009.

7 BagIt specification, <http://www.digitalpreservation.gov/partners/resources/tools/index.html#b> , accès en date du 19 mai 2009.

8 Django Web Framework, <http://www.djangoproject.com/> , accès en date du 19 mai 2009.

9 Consultez : Littman, Justin, "Actualized Preservation Threats: Practical Lessons from Chronicling America," *D-Lib Magazine*, 13:7/8 (juillet /août 2007).

<http://www.dlib.org/dlib/july07/littman/07littman.html> , accès en date du 20 mai 2009.

10 Consultez : Littman, Justin, "A Set of Transfer-Related Services," *D-Lib Magazine*, 15:1/2 (janvier/février 2009). <http://www.dlib.org/dlib/january09/littman/01littman.html> , accès en date du 20 mai 2009.

Merci à David Brunton, Ray Murray, et Deborah Thomas, de la Librairie du Congrès, pour leur contribution à cet article ...