TITRE : Revues savantes québécoises
SOUS-TITRE : des milliers d'articles en accès libre
AUTEURE :
- Nom, Prénom : Niemann, Tanja
- Coordonnées : tanja.niemann@umontreal.ca
- Titre : Coordonnatrice à l'édition numérique
- Lieu de travail : Érudit

AUTEURE :
- Nom, Prénom : Paquin, Émilie
- Coordonnées : emilie.paquin@umontreal.ca
- Titre : Analyste en gestion de l'information numérique
- Lieu de travail : Érudit

RÉSUMÉ :
Depuis 2007, Érudit a réalisé un vaste projet de numérisation rétrospective qui a permis la mise en ligne de plus de 30 000 articles de revues savantes québécoises. Cet article présente les différentes étapes de la chaîne de production éditoriale élaborée par Érudit et souligne au passage les défis de la numérisation rétrospective.

DATE : 22 juin 2009

Revues savantes québécoises : des milliers d'articles en accès libre

1. Numérisation rétrospective et patrimoine intellectuel

Nul besoin désormais de prendre l'avion ni de parcourir des kilomètres pour feuilleter sur les rayons d'une bibliothèque québécoise un numéro de la *Revue d'histoire de l'Amérique française*, fondée en 1947 par le chanoine Lionel Groulx. Quelques clics de souris suffisent à faire apparaître à l'écran les textes recherchés, extirpés de leur époque défunte par les activités de numérisation du consortium interuniversitaire Érudit.

La numérisation rétrospective effectuée par Érudit a permis, au cours des dernières années, la constitution d'une riche collection de revues universitaires québécoises. Des téléthéâtres d'Hubert Aquin, des extraits de romans en cours d'écriture d'Anne Hébert et des textes de Fernand Dumont ont depuis trouvé leur place au cœur de la sphère numérique mondialisée. L'édition, telle que pratiquée chez Érudit, repose sur des méthodes de production qui assurent l'accessibilité immédiate des publications aussi bien que leur préservation à long terme. Bien plus complexe qu'un simple transfert de support, l'édition numérique rétrospective pose des défis particuliers qui feront l'objet de cet article.

2. La numérisation rétrospective chez Érudit

La numérisation rétrospective n'est pas une pratique nouvelle chez Érudit. En 1999 déjà, Érudit numérisait les collections rétrospectives des revues *Meta* et *Sociologie et sociétés*. C'est toutefois une subvention du Ministère du Développement économique, de l'Innovation et de l'Exportation (MDEIE), obtenue en 2007, qui a permis de décupler les mises en chantier et d'accélérer la cadence des diffusions. Insuffisante pour assurer le traitement de l'entièreté des collections visées par le projet, la subvention du MDEIE a malgré tout donné un sérieux coup de pouce aux revues intéressées par le projet. Grâce aux efforts qu'elles ont déployés pour dénicher des fonds complémentaires, 21 des 26 revues concernées ont fait numériser la totalité de leur collection, ce que le financement initialement prévu n'aurait pas permis. Au printemps 2009, au terme du projet, ce sont plus de 30 000 articles qui ont été diffusés en accès libre. La numérisation et le balisage XML ont été réalisés par les deux pôles de production d'Érudit, le Centre d'édition numérique de l'Université de Montréal et la bibliothèque de l'Université Laval. Cela porte à 34 le nombre de revues diffusées dans leur intégralité sur la plateforme Érudit qui en compte à l'heure actuelle 79.

Les retombées de cette diffusion massive sont considérables. Non seulement ont-elles redonné aux Québécois un accès privilégié à des publications qui font désormais partie de leurs patrimoines culturel et scientifique, mais elles ont également assuré au réseau des universités québécoises une plus grande disponibilité des ressources documentaires. La revue de traductologie *TTR : traduction, terminologie, rédaction* est ainsi accessible dans son intégralité sur l'ensemble du territoire, ce qui n'était pas le cas auparavant.

Les statistiques de consultation révèlent éloquemment le succès rencontré par ce projet. Le nombre de visiteurs ayant consulté de la revue *Études littéraires* a en effet doublé à la suite de la mise en ligne de numéros rétrospectifs alors que celui de la revue *Voix et images* a été multiplié par six. Enfin, la visibilité d'Érudit, qui reçoit plus de 2 713 000 visiteurs par année avec plus de 11 985 500 pages consultées, en provenance du Canada, mais aussi des États-Unis, de l'Europe, de l'Amérique du Sud et de l'Afrique du Nord, assure aux publications

numérisées une présence appréciable sur la scène internationale et contribue assurément à faire connaître les résultats de la recherche publiés au Québec.

3. Étapes du processus de numérisation rétrospective
La réalisation d'un projet de numérisation rétrospective chez Érudit comporte 5 grandes étapes :

Constitution de la collection et gestion des droits d'auteur
⇓
Numérisation
⇓
Traitement et contrôle de qualité
⇓
Diffusion et promotion
⇓
Gestion des infrastructures et préservation

3.1 Constitution de la collection à numériser
Première étape de tout projet de numérisation rétrospective, la constitution de la collection est à la charge de la direction de la revue qui doit retracer les exemplaires à traiter et s'assurer de leur acheminement chez Érudit. Cette étape essentielle s'avère dans certains cas plus compliquée que prévu, lorsque des numéros de collections anciennes manquent ou que les éditions originales ne sont plus disponibles. Des recherches sont alors engagées par les directions de revues. Ces démarches peuvent s'échelonner sur plusieurs semaines et il arrive qu'un partenariat soit nécessaire à la réalisation du projet. Lors de la constitution de la collection de la *Revue de la Société historique du Canada*, certains numéros demeuraient introuvables. Érudit a dû avoir recours à l'aide de la bibliothèque de l'Université du Nouveau-Brunswick qui possédait les numéros recherchés. Impossible dès lors de massicoter (1) les exemplaires et de les traiter sur place, à Montréal. Ce sont donc les collègues de la bibliothèque de l'Université qui ont généreusement effectué une partie du travail et permis la mise en ligne de cette revue sur Érudit. Érudit est maintenant le seul organisme à donner accès à l'entièreté de la collection de la *Revue de la Société historique du Canada.*

Parallèlement à cette recherche de titres, la revue doit déterminer la marche à suivre concernant les droits des articles qui seront diffusés. La direction de chacune des revues est en effet responsable de cet important volet puisque, au final, les fichiers produits n'appartiennent pas au consortium mais bien aux éditeurs. Lorsque l'éditeur n'a pas signé de contrat de cession de droits avec ses auteurs, la revue doit faire des démarches auprès des titulaires des droits afin de s'assurer qu'elle détient les autorisations nécessaires à la diffusion numérique. Si un auteur refuse de voir un de ses articles paraître sur Érudit, un message est affiché au sommaire du numéro indiquant une restriction d'accès en raison du droit d'auteur.

Les articles sont malgré tout numérisés et traités de sorte que, si la revue ou le titulaire de droits venaient à invalider l'interdiction de publication (en raison du passage de l'article dans le domaine public ou si le titulaire des droits change d'avis), l'article puisse être diffusé sans que des coûts supplémentaires ne soient engagés.

Une fois la collection reconstituée, chacun des volumes fait l'objet d'une analyse détaillée visant à identifier les articles à numériser. C'est alors qu'il est déterminé si les notes de la rédaction, les index des auteurs ou encore les liminaires seront à conserver ou à exclure. Les considérations qui guident cette analyse ne sont pas de nature archivistique. Les numéros sont donc segmentés de sorte que chacune des unités documentaires scientifiquement pertinentes soit isolée. En cela, le travail effectué par le consortium s'éloigne des lignes directrices sur lesquelles se fondent, par exemple, les projets de numérisation patrimoniale menés par Bibliothèque et Archives nationales du Québec. Cela explique que les éléments satellitaires, comme les publicités ou les notices nécrologiques, ne soient pas diffusés. Une copie de l'intégralité de la revue est toutefois réalisée et conservée afin que toute trace de l'objet physique d'origine ne soit pas perdue. Un chercheur qui voudrait réaliser une étude sur l'usage de la publicité dans l'édition savante au Québec pourrait disposer des copies numériques des revues recensées par la recherche. À partir des résultats de cette analyse, la numérisation et la reconnaissance optique des caractères peuvent enfin être entreprises.

3.2 Numérisation

Conformément aux normes de production et de conservation développées par Érudit, quatre formats de fichiers sont générés. Un fichier PDF de type image-texte, permettant l'utilisation de la fonction « copier – coller », des fichiers RTF et TXT enregistrés en UTF-8 (2), et un fichier TIFF (3) avec une résolution de 600 dpi pour l'archivage. La reconnaissance optique des caractères est réalisée à partir du fichier TIFF et les volumes sont massicotés afin de réduire le coût de traitement. Toutefois, il peut arriver qu'il faille numériser à plat une revue dont il ne reste qu'un ou deux exemplaires.

C'est grâce à la reconnaissance optique des caractères (ROC) que l'indexation des articles dans l'outil de recherche pourra être réalisée, aussi est-il absolument essentiel que le résultat de la ROC soit impeccable. Un mot déformé par la reconnaissance optique de caractères nuit directement à la performance du moteur de recherche. Plusieurs petits détails doivent être considérés lors de cette étape : le retrait des tirets de césures (métamor-phose pour métamorphose), l'ajustement des paramètres permettant la reconnaissance automatisée des différentes langues utilisées (l'alternance du français et de l'anglais dans les références bibliographiques par exemple), l'exclusion des retours de chariot, etc. On estime que plus de 90% des mots d'un texte seront correctement interprétés par les logiciels de ROC. Évidemment, la qualité initiale de l'impression et les polices utilisées influencent grandement ces résultats. Une vérification est donc nécessaire afin de corriger les erreurs de ROC les plus flagrantes, dans les titres, les résumés et les bibliographies essentiellement.

3.3 Traitement des articles et contrôle de qualité

Les articles numérisés sont par la suite convertis en XML grâce à la chaîne de production éditoriale développée par les équipes Érudit de Montréal et de Québec. Le balisage des publications selon le schéma Érudit Article permet de représenter finement tous les éléments des articles scientifiques. Également utilisé par le portail français Persée (www.persee.fr) et par l'Electronic Text Centre (http://www.lib.unb.ca/Texts/) de l'Université du Nouveau-

Brunswick, le schéma Érudit Article constitue, véritablement, l'épine dorsale de l'entreprise éditoriale d'Érudit (4). Cette opération de balisage permet la création de métadonnées et l'identification des éléments sémantiques du texte. Les titres, auteurs, résumés, mots-clés ainsi que les bibliographies sont balisés. Toutefois, contrairement à la production des numéros courants, le corps des articles ne fait pas l'objet d'une analyse détaillée. L'insertion d'une balise en marque le début et la fin, ce qui est suffisant pour la recherche en texte intégral. Ce traitement minimal réduit considérablement les coûts de production sans nuire à la diffusion des articles pour chacun desquels un tiré à part en PDF est disponible. Néanmoins, un document XML de qualité est produit pour la recherche en texte intégral et tous les autres services disponibles dans Érudit.

Le travail éditorial effectué chez Érudit vise à opérer la translation d'une forme de représentation à une autre. Une double injonction guide l'entreprise : l'uniformité du traitement de la collection et la reproduction fidèle des documents traités. Or, les disparités dans une collection s'échelonnant sur plusieurs décennies peuvent être nombreuses. L'évolution des conventions éditoriales et les changements nombreux de directions qui ponctuent l'existence des publications universitaires complexifient le processus d'édition. Les incohérences typographiques ou éditoriales qui en découlent surgissent lors de la consolidation numérique de la collection, mais ces « erreurs » éditoriales ne peuvent être corrigées impunément puisque l'authenticité des publications est un principe cardinal chez Érudit. Une réflexion est donc être menée qui doit en outre prendre en considération les possibilités techniques de la plateforme de diffusion – et ses inévitables limitations. Ce qui pouvait passer pour une simple opération de transfert de support devient ainsi bien souvent une entreprise beaucoup plus complexe. Voilà pourquoi le contrôle de qualité qui chapeaute la production éditoriale revêt une importance aussi grande. Une validation éditoriale est effectuée par la coordonnatrice de l'édition numérique et par la direction de la revue avant toute diffusion.

3.4 Diffusion et promotion des nouvelles collections

Les articles sont diffusés en accès libre en raison du principe de la barrière mobile qui ne restreint l'accès des internautes qu'aux articles publiés au cours des deux dernières années. Les revues, financées par les fonds publics et numérisées grâce à des subventions gouvernementales, sont, de cette façon, accessibles aux universitaires aussi bien qu'au grand public. Cette politique favorisant le libre accès au patrimoine intellectuel diffère grandement des pratiques des éditeurs commerciaux qui restreignent l'accès des collections numérisées aux abonnés. La diffusion libre de ces collections décloisonne les résultats de la recherche universitaire et contribue, en cela, à une véritable démocratisation des savoirs, condition essentielle à l'exercice d'une citoyenneté éclairée.

Malgré les apparences, la mise en ligne des collections ne signifie pas la fin du travail effectué par Érudit, qui s'engage à promouvoir de façon continue les fonds numérisés. Grâce à des ententes de collaboration, les articles sont indexés dans des bases de données québécoises et internationales telles que Repère, ABC-Clio, FRANCIS et Google Scholar. La revue *Santé mentale au Québec* est même indexée par le prestigieux organisme américain PubMed. Les fonds sont également mutualisés en raison de partenariats établis avec des plateformes de diffusion des résultats de la recherche canadiennes et française, parmi lesquels figurent les Presses du CNRC (http://pubs.nrc-cnrc.gc.ca/rp-ps/journals.jsp?lang=fra), Synergies (www.synergiescanada.org) et Persée.

Érudit utilise différents canaux pour faire connaître son fonds, parmi lesquels les fils RSS, une liste de diffusion à l'intention des bibliothécaires (5), le Wiki Érudit (6), une page Facebook (7), et il assure, enfin, une représentation lors de congrès, de colloques et d'événements professionnels ou scientifiques. Des notices MARC produites par l'Université Laval sont fournies gratuitement aux institutions documentaires intéressées.

3.5 Gestion des infrastructures et préservation

Le passage du papier au numérique exige de la part des éditeurs un suivi continu des publications produites. Des problèmes de serveurs, un changement de logiciel, l'apparition de nouveaux outils informatiques, ou la manipulation malencontreuse de données peuvent altérer les documents diffusés, voire les rendre inaccessibles. Voilà pourquoi Érudit assure un accès pérenne aux revues et documents diffusés par l'utilisation de standards internationalement reconnus et en accès libre (XML, XHTML, TIFF). Une assistance aux utilisateurs individuels ou institutionnels aux prises avec des problèmes d'accès est aussi offerte de même qu'un système d'identifiants uniques. Ce dernier point est central, notamment pour les revues qui n'existent que sous forme numérique (8).

Enfin, la participation d'Érudit à la plateforme pancanadienne Synergies donne une assurance supplémentaire en regard de la préservation des documents puisque la plateforme souscrira à diverses initiatives de pointe en ce domaine, dont le format OAIS et l'initiative LOCKSS (9). De même, la collaboration d'Érudit avec Bibliothèque et Archives Canada dans le cadre d'un projet pilote visant à créer une structure pour le dépôt légal numérique permettra la conservation d'une copie de chacun des fichiers numériques produits. Le dépôt légal des publications numériques se fera également par Érudit, au nom des éditeurs, lorsqu'un tel service sera disponible à Bibliothèque et Archives nationales du Québec.

4. Quel avenir pour la numérisation rétrospective?

À l'heure actuelle, au moment de compléter la diffusion des revues ayant bénéficié de la subvention du MDEIE, nombreuses sont les revues à s'être montrées intéressées par une éventuelle publication rétrospective de leur fonds. Ainsi, loin de clore un chapitre, il semble bien que la fin de ce programme ouvre sur de nouvelles collaborations. Toutefois, malgré l'intérêt croissant de la communauté des chercheurs et des éditeurs, un projet de numérisation rétrospective reste une œuvre relativement coûteuse. Les revues savantes du Québec sont publiées par des organismes sans but lucratif, qu'il s'agisse d'associations, de presses ou de départements d'universités, et elles ne disposent généralement pas des fonds nécessaires pour mener à bien ces entreprises.

Dans ce contexte, il apparaît pertinent d'établir au Québec une politique de numérisation des patrimoines culturels et scientifiques. Alors que des entreprises privées et des consortiums commerciaux monnayent l'accès aux collections rétrospectives qu'ils numérisent, cette problématique du financement des projets rétrospectifs prend une importance nouvelle puisque c'est, véritablement, du contrôle de l'accessibilité aux contenus numérisés dont il est question. Il est juste d'affirmer que, par l'action d'Érudit dans ce chantier de numérisation, des économies de frais d'abonnement sont réalisées tout en assurant une diffusion des revues québécoises sans précédent dans notre histoire. Cette opération, rendue possible grâce au soutien du MDEIE, du FQRSC et des universités membres du consortium,

bénéficie à l'ensemble des Québécois et contribue à bâtir, dans une perspective de service public, l'infrastructure numérique en réseau des bibliothèques du Québec.

Encadré
Revues dont la collection a fait l'objet d'une numérisation rétrospective :
L'Actualité économique
Anthropologie et Sociétés
Cahiers de géographie du Québec
Cahiers québécois de démographie
Cinémas
Criminologie
Études françaises
Études internationales
Études littéraires
Géographique physique et Quaternaire
Horizons philosophiques
Journal of the Canadian Historical Association / Revue de la Société historique du Canada
Laval théologique et philosophique
Laval théologique et philosophique
Lien social et Politiques
Meta
Nouvelles perspectives en sciences sociales
Nouvelles pratiques sociales
Philosophiques
Politique et Sociétés
Phytoprotection
Protée
Recherches féministes
Recherches sociographiques
Reflets : revue d'intervention sociale et communautaire
Relations industrielles / Industrial Relations
Revue d'histoire de l'Amérique française
Revue québécoise de linguistique
Revue des sciences de l'eau / Journal of Water Science
Revue des sciences de l'éducation
Romanticism on the Net
Santé mentale au Québec
Scientia Canadensis
Service social
Sociologies et sociétés
Tangence
Théologiques
TTR
Voix et Images

Notes
(1) Selon le Petit Robert, massicoter signifie « rogner, couper (le papier) au massicot », c'est-à-dire découper le dos de la revue pour que les feuilles ne soient plus reliées.

Le Petit Robert. 2000. Massicoter : p. 1531.

(2) « UTF-8 (UCS transformation format 8 bits) est un format de codage de caractères défini pour les caractères Unicode. »
Wikipédia. 2009. *UTF-8* [http://fr.wikipedia.org/wiki/UTF-8] (consultée le 20 juin 2009)

(3) « TIFF est un format extrêmement flexible. Il est notoirement connu pour permettre l'enregistrement des données multi-octets au format big endian ou little endian. Il permet d'utiliser de nombreux types de compression, avec ou sans perte de données : brut, PackBits, LZW, CCITT Fax 3 et 4, JPEG. Il supporte de nombreux codages des pixels, de 1 à 64 bits par pixel, signé ou non, ainsi que les formats en virgule flottante 32 et 64 bits définis par l'IEEE. Il supporte de nombreux espaces colorimétriques : noir et blanc, monochrome, palette de couleurs (de toute taille), RVB, YCbCr, CMJN, CIELab. Il supporte de nombreuses informations additionnelles sur les couleurs utiles à la calibration colorimétrique : correction gamma, etc. Il supporte le stockage d'image par bloc, et aussi de multiples images par fichier, des images alternatives en basse résolution, des annotations sous forme de courbes et de texte, etc. »
Wikipédia. 2009. *TIFF* [http://fr.wikipedia.org/wiki/Tagged_Image_File_Format] (consultée le 20 juin 2009)

(4) Pour en savoir davantage sur le schéma Érudit Article, voir la documentation sur le Wiki Érudit : https://cen.umontreal.ca/pages/viewpage.action?pageId=5734590.

(5) Pour s'abonner à la liste de diffusion : http://www.listes.umontreal.ca/wws/subrequest/erudit-bibliolib.

(6) Wiki Érudit : https://cen.umontreal.ca/display/Erudit/Bienvenue

(7) Pour devenir fan d'Érudit : http://www.facebook.com/pages/edit/?id=141469165320#/pages/Erudit/141469165320

(8) En octobre 2003, la revue *Science* publiait en effet les résultats d'une enquête qui révélait qu'après trois mois de publication seulement, 3,8% des ressources internet de trois revues majeures en médecine n'étaient plus actives. Ce pourcentage grimpait à 10% après 15 mois et à 27% 22 mois plus tard. Plus désastreuse encore, une autre étude, publiée en 2007 par *The Serial Librarian,* démontrait que la moitié des ressources disponibles sur Internet dans trois des plus importantes revues en communication aux États-Unis n'étaient plus disponibles trois ans après leur parution.
Moghaddam, Golnessa Galyani. 2007. "Archiving Challenges of Scholarly Electronic Journals : How do Publishers Manage Them", Serials Review, vol. 33 n°2, pp. 81-90.

(9) Lots of Copies Keep Stuff Safe (www.lockss.org).