arbido

2008/2 Die Wirtschaft der neuen Medien

Fortun Virginie,

La mise en ligne des archives du Journal de Genève

Un quotidien est toujours le témoin de son époque. En cela, la numérisation de la collection complète du *Journal de Genève* est apparue comme une évidence. L'objectif était d'assurer des conditions de conservation à long terme à ce fabuleux patrimoine écrit et de faciliter ses possibilités de consultation.

Le *Journal de Genève* (1826–1998) représente une source d'information inestimable relatant la vie économique, politique, culturelle et sociale de notre pays sur plusieurs siècles. Toutefois cette source est largement sous-exploitée compte tenu de ses conditions d'accès et de recherche très contraignantes. Pour des raisons historiques et à son initiative, *Le Temps* a entrepris la numérisation et la mise en ligne de ce prestigieux titre. Les discussions préalables ont été initiées en totale collaboration avec la société anonyme du *Journal de Genève* afin de régler les aspects juridiques, financiers et techniques.

Un véritable partenariat public-privé

Après consultation et en raison de l'importance historique et de l'ancienneté du titre, la Bibliothèque nationale suisse a décidé de participer au projet. En tant que bibliothèque patrimoniale du lieu de parution du titre et dépositaire légal, la Bibliothèque de Genève a également confirmé son implication.

L'établissement d'un partenariat public-privé a permis de partager des ressources non seulement financières mais aussi humaines, réunissant ainsi l'ensemble des compétences nécessaires à une mise en œuvre professionnelle. Les prestations ne pouvant être assumées par les partenaires du projet, à savoir le *scanning* et la reconnaissance optique de caractères, ont été confiées à des tiers selon une procédure d'appel d'offres.

Le financement du projet a été assuré d'une part par le soutien de la Bibliothèque nationale et de la Bibliothèque de Genève ainsi que par la recherche de fonds externes opérée auprès de mécènes et de sponsors d'autre part. Le coût complet de numérisation d'une page à partir des originaux papiers avoisine en effet CHF 1.–.

Devenir cherchable donc trouvable

In fine, l'ambition commune a toujours été de proposer une recherche plein texte sur l'intégralité du corpus afin d'autoriser une recherche avancée performante, selon les désirs et les besoins de chaque utilisateur. Pour ce faire, le projet s'est articulé autour de trois étapes principales.

– La préparation. Cette étape a été consacrée à la définition des besoins, l'identification des standards de numérisation, la sélection des volumes, l'établissement des conventions unissant les partenaires ainsi que la recherche de financement. La numérisation étant une noble cause mais pas forcément prioritaire en regard de projets caritatifs ou liés à la santé, cette recherche a été particulièrement chronophage.

La faible expérience helvétique en matière de numérisation nous a conduits à étudier les projets européens afin de rechercher les meilleures pratiques, de s'assurer de la pertinence de nos choix et de notre conformité avec les standards les plus usités.

– Le *scanning.* C'est la phase du projet qui a demandé le plus de temps et d'attention. Tout a commencé par la rédaction de l'appel d'offres et la sélection du prestataire. La fragilité des volumes imposait que la numérisation n'entraîne aucune contrainte mécanique sur les originaux et que l'exposition au rayonnement ultravio- let et aux infrarouges soit réduite au minimum. Le transport des volumes hors de Suisse n'a, pour ces raisons, jamais été envisagé.

Une fois la convention de nommage établie, chaque page a été numérisée avec une définition de 300 DPI, sans interpolation, et traitée en 256 niveaux de gris (bits). Une page correspondant à un fichier. Pour chaque page, la digitalisation a permis de créer un fichier image (tiff 6.0, mono page non compressé), un fichier PDF et un fichier texte puisqu'une première reconnaissance optique de caractères a été opérée à la volée lors du *scanning.* Véritable projet dans le projet, la gestion de ce patrimoine numérique a également été un enjeu de taille. En effet, au-delà de la conservation des fichiers sources, la question de la pérennité des standards et surtout de la disponibilité des moyens de lecture de ces données à très long terme par les générations futures est une véritable question. La capacité de stockage et la manipulation de ces innombrables fichiers ont donc requis un soin tout particulier. Au-delà de facteurs exogènes, la longue durée de cette étape est imputable à deux éléments principaux. Tout d'abord le débrochage de la moitié de la collection qui n'avait pas du tout été prévu lors de la définition du planning de production et qu'il a fallu opérer en cours de projet pour faciliter l'adaptation des volumes au *scan.* Puis, les tests de qualité, de cohérence et d'exhaustivité qui ont été menés de manière aléatoire sur le million de fichiers numériques nouvellement créés.

– La reconnaissance optique de caractères (OCR) et la mise en ligne. Comme pour le *scanning* la première étape a consisté en la sélection du prestataire. La plus grande mobilité des données, puisque devenues électroniques, autorisant une recherche de prestataire à l'échelle internationale.

Le transfert des données, la définition de la politique de segmentation, la reconnaissance des différents éléments structurels puis l'exportation des métadonnées comme objets XML basés sur les schémas reconnus METS et ALTO et le contrôle qualité ont été les principales étapes de l'OCR. L'accessibilité de l'information a guidé notre démarche. L'application de consultation devait donc pouvoir supporter plusieurs titres et plusieurs langues. L'accès aux contenus est possible par édition, par les unes, la recherche avancée supporte les critères booléens et permet d'isoler l'information selon sa localisation (articles, annonces, publicités). L'article peut toujours être visualisé dans le contexte de la page complète. L'infras- tructure technologique a été dimensionnée pour assurer la rapidité dans l'affichage des résultats, la sécurité et la continuité des données.

La version numérisée du journal possède d'indéniables avantages pour les utilisateurs, qu'ils soient chercheurs, étudiants, bibliothécaires, journalistes ou simples citoyens. La consultation simultanée sur place ou à distance, l'optimisation des recherches (par mot clé, type de contenus, date ou édition), le fait que la dégradation des supports numérisés soit indépendante du nombre de lecture ou encore que la qualité demeure constante entre les duplications et les transmissions en sont quelques illustrations. La pérennité du titre est donc désormais assurée sur le plan technique. L'accroissement de sa consultation, parce que plus aisée, va également contribuer à lui donner une deuxième vie.

Ampleur de la collection complète du *Journal de Genève*

– 172 années de parution (1826–1998)

– Rythme hebdomadaire devenu quotidien dès 1850 – 600 000 pages

– 30 mètres linéaires

– Sélection des volumes les mieux conservés à partir de trois collections

– 6 formats différents sur toute la durée de parution

– 13 TO de données numérisées

– 24 mois de réalisation

– Collection complète accessible via le web en automne 2008

Partager les connaissances

L'expérience acquise lors de ce projet pionnier en Suisse romande va permettre aux prochains porteurs de projet d'éviter les écueils d'une première réalisation, de créer des synergies, d'allouer efficacement les ressources financières et de gagner du temps. En effet, la digitalisation du *Journal de Genève* a donné l'impulsion à la conclusion d'un accord entre la Bibliothèque nationale, les bibliothèques cantonales de Suisse romande, le réseau des bibliothèques de Suisse occidentale et Presse Suisse. Cet accord cadre intitulé *Un patrimoine en ligne* est un catalogue de recommandations pour la mise sur pied de collaborations ponctuelles entre tel éditeur et telle bibliothèque, concernant un ou plusieurs titres de journaux.

Dans cette même logique de partage des connaissances, l'accès au *Journal de Genève* se veut libre pour que le plus grand nombre puisse pleinement bénéficier de cette source d'information pluricentenaire.

Dans un contexte d'abondance de l'information, les éditeurs doivent à présent évoluer dans une nouvelle ère: l'économie de l'attention. Parmi une offre pléthorique, la ressource rare devient en effet la disponibilité de l'utilisateur final, toujours davantage sollicité. Le défi ultime d'un éditeur est donc de proposer une information adaptée à la demande de chaque lecteur, qu'il s'agisse du sujet, du format, du véhicule de transmission et de la temporalité. La numérisation du *Journal de Genève* s'inscrit dans la stratégie numérique du *Temps,* qui vise à ce que l'ensemble de ses contenus et le patrimoine dont il est dépositaire soient le plus efficacement disponibles. Il s'est donc écoulé 182 ans entre la création des contenus du premier numéro du Journal et leur mise à disposition sur le web. Le cycle est ainsi complet et sera prochainement soumis à l'attention des utilisateurs. Nul doute que le *Journal de Genève* sera rapidement rejoint par d'autres titres romands.

Fondé en mars 1998, *Le Temps* est le quotidien de référence de la Suisse romande et francophone. «Quality paper» aux axes d'expertise reconnus en politique suisse et internationale, en économie et finance et en culture, *Le Temps* privilégie la rigueur dans l'information, la compétence dans l'analyse et la pluralité dans l'expression des opinions. Son lectorat s'élève à 131 000 personnes. (Source REMP MACH Basic 2008-1)

Les enseignements

a) Le financement est l’enjeu majeur

b) La sélection des volumes est fondamentale

c) Les aspects logistiques ne sont pas à sous-estimer

d) Le contrôle qualité est chronophage MAIS indispensable

e) La gestion de projet doit intégrer le long terme

Virginie Fortun

Marketing Manager, LE TEMPS SA

Abstract

Deutsch
Das Journal de Genève (1826–1998) ist eine Informationsquelle von unschätzbarem Wert. *Le Temps* hat beschlossen, diese Quelle online zur Verfügung zu stellen. Angesichts des hohen historischen Werts des Archivs hat sich die Schweizerische Nationalbibliothek dazu entschlossen, sich am Projekt zu beteiligen. Ebenfalls beteiligt hat sich die Bibliothèque de Genève, welche gleichzeitig die gesetzliche «Nachlassverwalterin» des Zeitungstitels ist. Finanziert wurde das Projekt von der Nationalbibliothek, der Bibliothèque de Genève sowie von privaten Mäzenen und Sponsoren. Es wurde errechnet, dass die Digitalisierung einer Zeitungsseite auf ca. CHF 1.– zu stehen kommt. Es war von Beginn weg das Ziel, eine Volltextsuche für den gesamten Korpus anzubieten. Um dieses hochgesteckte Ziel erreichen zu können, wurde das Projekt in drei Etappen aufgeteilt:

- **Vorbereitungsphase. In dieser Phase wurden die Bedürfnisse abgeklärt, Standards für die Digitalisierung gesetzt, erste Auswahlen getroffen, rechtliche Abklärungen vorgenommen und die Finanzierung organisiert.**

- **Scanning. Der Auftrag wurde ausgeschrieben. Die z.T fragilen Originale mussten entsprechend vorsichtig behandelt werden. Als Auflösung wurden 300 dpi bei 256 Graustufen gewählt. Eine Seite entsprach einer Datei. Jede Seite wurde in den Formaten TIFF, PDF und DOC abgespeichert. Das Speichern der umfangreichen Datenmenge und die Abklärungen zur Frage, wie die Daten langfristig gespeichert werden müssen, damit sie auch von nachfolgenden Generationen genutzt werden können, bereitete etliches Kopfzerbrechen.**

- **Texterkennung (OCR) und Aufschaltung. Wie schon beim Scanning wurde nach einer Ausschreibung ein externer Anbieter mit der Aufgabe betraut.**

  Oberstes Ziel war und blieb die Zugänglichkeit der Informationen. Die Suchmaske ermöglicht es, nach mehreren Titeln und in mehreren Sprachen zu suchen. Von der technischen Seite her wurde darauf geachtet, dass die Geschwindigkeit bei der Anzeige der Resultate, die Sicherheit und die Kontinuität der Daten permanent gewährleistet sind. Die digitale Version der Zeitung bietet den Nutzerinnen und Nutzern unzweifelhaft zahlreiche Vorteile. Die zahlreichen Suchoptionen und die Möglichkeit, von überall her auf die Daten zuzugreifen, werden dem «Journal» voraussichtlich ein zweites Leben einhauchen.

  Von den Erfahrungen, die im Rahmen dieses Pionierprojekts in der Romandie gesammelt wurden, werden andere Projektverantwortliche, die Ähnliches im Sinn

haben, profitieren können. Die Digitalisierung des Journal de Genève führte letztlich zur Verabschiedung von Rahmenempfehlungen (Titel: *«Un patrimoine en ligne»*) für die Digitalisierung von Zeitungen in der Westschweiz. Mitgetragen haben diese Ausarbeitung von Rahmenempfehlungen die Nationalbibliothek, die Westschweizer Kantonsbibliotheken, das Réseau des bibliothèques de Suisse occidentale (RERO) und Presse Suisse. Es handelt sich dabei um einen Katalog mit Empfehlungen, wie die – punktuelle – Zusammenarbeit zwischen einem Herausgeber und einer Bibliothek im Hinblick auf die Digitalisierung von einem oder mehreren Zeitungstiteln gestaltet werden könnte.

Der Zugang zum digitalisierten Archiv des Journal de Genève ist übrigens frei. Auch hier war die Absicht, dass möglichst viele das Angebot nutzen und davon profitieren können. Ein paar Kennzahlen zum Archiv des Journal de Genève

- **172 Erscheinungsjahre (1826–1998)**
- **zu Beginn eine Wochenzeitung, ab 1850 Tageszeitung**
- **600 000 Seiten**
- **30 Regalmeter**
- **Über die gesamte Erscheinungsdauer wurde das Format sechsmal geändert.**
- **13 Terabytes digitalisierte Daten**
- **Projektdauer = 24 Monate**
- **Das komplette Archiv ist ab Herbst 2008 online.**