



**La qualité de la quantité : la numérisation de la presse à la Koninklijke Bibliotheek (Bibliothèque nationale des Pays-Bas)**

**Edwin Klijn**

Chef de projet « Base de la presse quotidienne numérisée »  
Koninklijke Bibliotheek  
La Haye, Pays-Bas

*Traducteur : Philippe Vallas*  
*Bibliothèque nationale de France*  
[Philippe.vallas@bnf.fr](mailto:Philippe.vallas@bnf.fr)

**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 et introduction :**

*En 2006, après qu'une subvention ait été octroyée par l'Organisation néerlandaise pour la recherche scientifique (NWO), un important projet de numérisation de la presse commença à être élaboré – la Base de la presse quotidienne numérisée ainsi qu'on l'appela. Bien que la Koninklijke Bibliotheek ait été engagée sans interruption dans des activités de numérisation depuis 1995, jusqu'il y a cinq ans elle manquait d'expérience pour produire rapidement un grand nombre de fichiers. En 2003 elle lança son premier projet de numérisation de masse, concernant les Documents parlementaires néerlandais (2,5 millions de pages). Ce fut un jalon déterminant dans les activités de numérisation de l'établissement : la numérisation « artisanale » de haute qualité, centrée sur des images visuellement attractives, était remplacée par une numérisation de masse centrée sur les textes, pour laquelle importait non seulement la qualité, mais aussi la quantité. Pour la Koninklijke Bibliotheek, le passage à la numérisation de masse a constitué un nouveau pas en avant, de bon sens : après des années d'apprentissage et d'expérimentation, le moment était venu d'avancer et de passer à la mise en pratique à grande échelle. Avec l'avènement de l'ère de la numérisation à la bibliothèque, des questions nouvelles et parfois inédites sont apparues, sur lesquelles il fallait immédiatement se pencher malgré le flot des obligations quotidiennes. La présente communication aborde certaines de ces questions, qui se sont faites jour au cours du projet de numérisation des journaux, lequel est maintenant presque à mi-chemin. Elle constitue une étude de cas sur l'impact organisationnel de la numérisation de masse dans un environnement de bibliothèque.*

## **Sélection et droits d'auteur**

Bien que portant sur 8 millions de pages de presse nationale, régionale, locale et coloniale néerlandaise, le projet ne concerne cependant qu'une proportion réduite (8 % environ) du total des journaux publiés dans le pays toutes époques confondues. Un comité scientifique consultatif, composé d'historiens, de journalistes, de linguistes et autres « utilisateurs intensifs », sélectionne les titres en fonction de leur intérêt pour la communauté des chercheurs. Beaucoup de ces titres n'étant pas conservés par la Koninklijke Bibliotheek, des institutions extérieures, nationales ou étrangères (Stockholm, Saint-Petersbourg, Londres, Oldenburg, Dresde, Paramaribo, le Vatican) sont sollicitées pour un prêt temporaire de leurs collections originales, ou pour fournir elles-mêmes les fichiers. Lorsque les journaux sont déjà disponibles sur microfilms ceux-ci font l'objet d'un contrôle de leur qualité et de leur complétude. Dans le principe, la numérisation des microfilms est préférée à celle des originaux sur papier, car moins coûteuse pour la prise de vues et nécessitant moins de travail pour compléter les collections. Les pertes de qualité d'image, et donc d'OCR, qui en résultent sont jugées acceptables – dans certaines limites – dans la mesure où le budget économisé en ne recherchant pas la plus haute qualité possible permet en compensation de numériser un plus grand nombre de pages.

Le projet a l'ambition de numériser des journaux jusqu'à l'année 1995. Par conséquent, pour presque tous les titres de presse du 20<sup>e</sup> siècle (estimés à 1,5 million de pages), la législation néerlandaise sur le droit d'auteur est applicable. Les détenteurs de droits peuvent être nombreux dans le cas des journaux : éditeurs, mais aussi journalistes indépendants, illustrateurs, photographes, etc. Selon la législation néerlandaise, une œuvre imprimée tombe dans le domaine public 70 ans après sa publication et 70 ans après la mort du dernier détenteur de droits. Par exemple, si un photographe qui a publié une photographie en 1925, à l'âge de 20 ans, meurt à l'âge de 70 ans, ce travail tombera dans le domaine public en 2045. Lorsqu'elle « re-publie » en ligne ces journaux, la Koninklijke Bibliotheek est légalement tenue de retrouver au préalable tous les ayants droit, ce qui est virtuellement impossible pour une aussi grande quantité de titres différents. C'est pourquoi elle a entrepris de négocier avec les instances représentatives des différentes catégories d'ayants droit pour trouver une solution globale. Dans le cas des indépendants, l'idée serait de mettre au point un système qui déchargerait les instances spécialisées de leur obligation légale de localiser les ayants droit, en échange d'une contribution financière consistant en un pourcentage spécifique du budget global du projet. Le projet de numérisation des journaux est considéré comme un test national. Le modèle économique qui résultera des discussions deviendra très vraisemblablement dans l'avenir la référence pour toutes les actions de numérisation menées par les institutions patrimoniales néerlandaises sur des collections originales non-tombées dans le domaine public.

## **Préparation matérielle**

Après avoir procédé à un appel d'offres européen, la bibliothèque a externalisé la numérisation, l'océrisation et la post-production chez un prestataire extérieur (CCS de Hambourg et son sous-traitant M&R de Kampen). Depuis juin 2008, environ 60 volumes sont transportés chaque semaine de la bibliothèque jusqu'à l'atelier de numérisation de Kampen, distant d'environ 120 km. A la Koninklijke Bibliotheek, une équipe composée actuellement d'environ 10 personnes est responsable de la préparation des documents à la numérisation. Au départ, la majorité d'entre eux était employée uniquement sur le projet de numérisation des

journaux, mais comme la bibliothèque a étendu ses activités de numérisation à d'autres types de documents imprimés (par exemple aux monographies, aux revues, etc), ils travaillent aussi à présent pour d'autres projets. Leur activité comprend essentiellement les petites réparations (afin d'obtenir une numérisation de qualité), l'insertion des métadonnées dans une base, et la préparation des transports. Spécifique au projet à l'origine, l'équipe de préparation des collections a été récemment insérée dans l'organigramme régulier de la bibliothèque. Elle fait maintenant officiellement partie du département responsable des collections originales de l'établissement, ce qui présente des avantages importants. L'expertise sur les documents physiques est intégrée dans la chaîne de numérisation. Les procédures de préparation, de conditionnement et de transport mises au point pour le projet de numérisation de la presse sont à présent réutilisées pour tous les projets et programmes de numérisation de la bibliothèque.

## **Numérisation**

Le planning du projet impose au prestataire de traiter 65 000 pages par semaine en moyenne. Ce chiffre inclut non seulement la prise de vue, mais aussi l'océrisation et la post-production. Pour les journaux, les métadonnées créées par l'équipe de préparation matérielle des documents sont insérées et complétées par d'autres métadonnées qui distinguent les différents articles présents dans une page, le « type » d'article (nouvelles, publicité, carnet, illustration avec sa légende), et par un titre modifié manuellement. Le prestataire a installé à Kampen une unité de production capable d'effectuer une numérisation globale puis la segmentation des journaux en articles. Automatiser autant de maillons que possible dans la chaîne de production est un travail nécessitant beaucoup de précision et de temps. Une petite erreur, un petit défaut de communication peuvent affecter un grand nombre de pages. Passer par de tels systèmes de production implique aussi, inévitablement, d'accepter des compromis. Par exemple, pour les fichiers PDF noir et blanc produits pour chaque numéro de journal, la bibliothèque a dû accepter que l'outil de réglage du contraste déclenchant la prise de vue soit incapable de traiter correctement les journaux comportant beaucoup de texte visible par transparence sur l'autre face des feuillets. Passer à des PDF en niveaux de gris était impossible car cela aurait multiplié au moins par trois le poids des fichiers.

## **Contrôle de la qualité**

Après leur production par le prestataire, les fichiers sont livrés chaque semaine à la bibliothèque sur des disques durs. Plusieurs contrôles sont effectués sur les lots, certains d'entre eux entièrement automatiques (p. ex. le contrôle de la validité et de la bonne constitution des fichiers XML), d'autres effectués « manuellement » par l'équipe de préparation des collections (p. ex. le contrôle de la segmentation du texte en articles) et les gestionnaires de contrôle qualité (p. ex. la définition et la gamme de gris des images). Le contrôle-qualité comprend beaucoup d'opérations, car il faut s'assurer de la qualité des fichiers livrés. Sur le premier million de pages, environ 10 % de l'ensemble des fichiers reçus ont été rejetés. Ces chiffres montrent que même sur les chaînes de production hautement automatisées le risque d'erreur est considérable, et souvent sous-estimé.

## **Accès, présentation, préservation numérique**

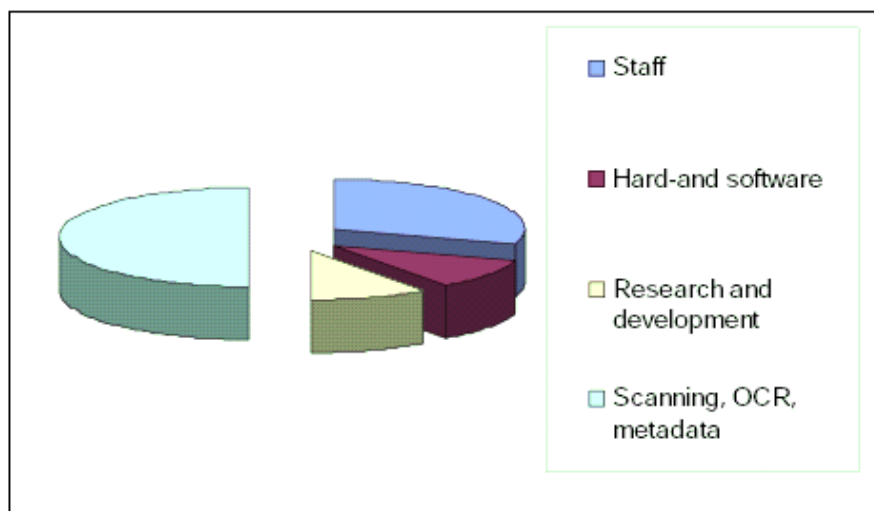
Une fois qu'ils ont passé le contrôle-qualité, les fichiers sont intégrés dans l'environnement informatique de diffusion et d'archivage de la bibliothèque. On va devoir créer un index pour tous les fichiers d'articles, qui comprendra quelque 64 millions de fichiers textes. Avec cette quantité, la grosse difficulté est de concevoir un système de recherche performant. Les statistiques obtenues actuellement à partir du site-pilote de consultation de la presse laissent à penser que la Base de la presse quotidienne numérisée va générer un plus grand nombre de consultations que toute autre application de la bibliothèque déjà opérationnelle. On a développé pour les besoins de ce projet un visualiseur adapté aux fichiers en format de consultation JPEG2000, conçu comme un service pouvant être réutilisé pour les autres projets de numérisation.

Sur le site internet les usagers pourront effectuer des recherches plein-texte sur approximativement 25 milliards de mots. La recherche avancée sera également possible par période chronologique spécifique, titre, lieu de publication, type d'article, et/ou aire de diffusion. La qualité de l'OCR devrait varier d'un titre à l'autre. Afin d'améliorer l'accès à des corpus de textes aussi énormes, la Koninklijke Bibliotheek participe à un projet européen nommé IMPACT (<http://www.impact-project.eu/>). Le projet de numérisation de la presse permettra de tester les outils et l'expertise obtenus grâce à IMPACT.

Les fichiers-images maîtres seront avec certaines de leurs métadonnées conservés dans l'E-Dépôt, le système de préservation numérique de la bibliothèque. Le projet de numérisation de la presse, qui n'occupera qu'une petite partie du total, devrait générer environ 120 To de données à conserver de façon pérenne. Les coûts de stockage annuels d'un To dans l'entrepôt numérique sont estimés actuellement à 8 500 €. Il est évident que la maintenance de tous les documents numériques aura un impact croissant sur la bibliothèque au cours des prochaines années.

## **Le futur**

Il y a beaucoup de marge d'amélioration dans la chaîne de production actuelle. Le coût global par page (1,50 € tout compris) est relativement élevé. Cela n'est pas dû aux coûts de numérisation, mais surtout aux exigences élevées du projet quant à la précision et à la qualité des métadonnées. De plus, en raison des efforts importants consacrés à la sélection, à la préparation des documents et au contrôle de la qualité, les coûts de personnel représentent plus du quart du budget global ( cf. diagramme n° 1). En rendant plus fluide la chaîne de production et en réutilisant pour d'autres projets les connaissances et l'expérience acquises, la bibliothèque s'efforce de gagner en efficacité dans ses activités de numérisation.



*Diagramme 1 : répartition des coûts dans le projet Base de la presse quotidienne numérisée [Personnel; Logiciel et matériel informatique; Recherche et développement; Numérisation, ROC, métadonnées]*

Jusqu'en 2013, la Koninklijke Bibliotheek est engagée dans de grands changements. L'objectif est de rendre accessibles sous forme numérique environ 10 % de l'ensemble des documents imprimés publiés aux Pays-Bas toutes époques confondues. Il reste de nombreux obstacles à surmonter pour atteindre ce but. En première ligne de cette évolution, le projet Base de la presse quotidienne numérisée permet à la bibliothèque d'acquérir une expérience pratique des nombreuses facettes de la numérisation de masse. Espérons que dans le futur on rira bien fort à l'idée qu'en 2009, pour un simple total de 8 millions de pages, on parlait sérieusement de numérisation DE MASSE.

Edwin Klijn  
 Chef de projet « Base de la presse quotidienne numérisée »  
 Koninklijke Bibliotheek  
[Edwin.klijn@kb.nl](mailto:Edwin.klijn@kb.nl)  
 Site du projet : <http://www.kb.nl/projectdagbladen/>

*Juillet 2009*