



Microfilmage en vue de la numérisation et de la reconnaissance optique de caractères

Supplément aux Directives

Introduction

Les [Directives pour la réalisation de microfilms de conservation des journaux](#) furent publiées par IFLA en 1996 puis montées sur IFLANET.

La Section IFLA des Journaux (anciennement Table ronde pour les Journaux) a prévu de publier un supplément aux *Directives*. Elle est en effet consciente de la nécessité d'assurer à l'avenir le meilleur accès possible aux textes reproduits sur microfilm au cas où ces films feraient l'objet d'une numérisation.

Ce travail a récemment connu un nouvel élan significatif en raison de :

- la croissance rapide d'Internet qui renforce la communication de textes parus dans la presse courante ou ancienne
- l'amélioration importante des possibilités des machines disponibles qui, avec les ensembles de logiciels performants, permettent d'obtenir des résultats de grande qualité en numérisant rapidement les textes
- la baisse des frais de stockage des données numérisées et la capacité d'en stocker des quantités beaucoup plus importantes que par le passé

Au moment où les développements techniques changent notre perception de ce qui est possible, année après année, la Section des Journaux considère qu'il est approprié de publier un document d'orientation intitulé : *Microfilmage en vue de la numérisation et de la reconnaissance optique de caractères*. Pour de nombreuses bibliothèques le microfilm demeure le support préféré pour l'archivage de textes. Optimiser l'accès à des microfilms peut être le meilleur moyen à l'avenir de rendre universel l'accès à d'énormes masses de journaux dont certains sont déjà conservés sur microfilm alors que d'autres sont encore à microfilmer.

La numérisation a permis d'améliorer la mise à disposition des journaux en ajoutant à l'accès aux textes des moyens de recherche performants. En raison de leur consultation intense les bibliothèques consacrent une part importante de leurs programmes de numérisation aux journaux. Mais la numérisation a des exigences particulières quant à la qualité du microfilmage.

La numérisation de journaux qui se sert du microfilm comme support intermédiaire nécessite un niveau de qualité du microfilm plus élevé. Pour le projet nordique TIDEN la Bibliothèque de l'Université d'Helsinki et la Bibliothèque Royale de Suède ont testé l'impact de la qualité du microfilm sur la numérisation et la reconnaissance optique de caractères. Etant donné que les résultats de ces tests sont susceptibles d'intéresser

un public plus vaste, les participants au projet (auxquels se sont jointes la Bibliothèque Nationale de Norvège et la Bibliothèque d'Etat d'Aarhus au Danemark) ont contribué, pour le compte de la Section IFLA des Journaux, à l'élaboration de recommandations pour le *Microfilmage en vue de la numérisation et de la reconnaissance optique de caractères*.

Dans le cadre du projet nordique TIDEN ont été numérisés des journaux nordiques parus entre 1640 et 1900, principalement imprimés en caractères gothiques mais également en caractères romains. Le microfilm a été utilisé comme support intermédiaire. Outre les commentaires émis par les membres de l'ex-Table ronde pour les Journaux, il faut rappeler en particulier la collaboration de la Bibliothèque nationale de France et de la British Library. Tous les résultats réunis seront intégrés comme supplément aux *Directives pour la réalisation de microfilms de conservation des journaux* publiées précédemment par cette Table ronde.

Etendue de compétence

L'accès amélioré aux données stockées sur supports électroniques, ainsi que la pérennité du microfilm, offrent une bonne opportunité pour l'archivage et la communication. La durée de vie du microfilm avoisine 500 ans, ce qui assure la sauvegarde de textes transférés à partir d'originaux dont le papier s'effrite. La numérisation en facilite l'accès. Mais le microfilm est davantage qu'un support d'archivage. Il est aussi un intermédiaire efficace et économiquement rentable dans un programme de numérisation si cet aspect est bien pris en compte lors du processus de microfilmage.

Le document d'orientation a pour objectif de montrer comment on peut utiliser le microfilm comme plate-forme pour la numérisation ultérieure et pour permettre l'interrogation en texte intégral des collections numérisées. Etant donné que l'évolution est extrêmement rapide, il faut souligner que les suggestions publiées dans le document IFLA intitulé *Microfilmage de conservation en vue de la numérisation et de la reconnaissance optique de caractères* sont le reflet de la situation actuelle. De nouvelles techniques verront le jour.

Il est possible de combiner le microfilmage et la numérisation de plusieurs manières: microfilmer d'abord puis numériser le film ; ou bien, numériser d'abord puis conserver le contenu sur film en utilisant la méthode COM (computer output microfilm/microfilm de sortie d'ordinateur) ; enfin utiliser une caméra hybride pour microfilmer et numériser simultanément. Pour les journaux la première option, à savoir commencer par le microfilm, est de loin la plus utilisée. C'est aussi la méthode que nous allons présenter en détail ici. (Toutefois, au cas où la bibliothèque n'a pas encore fait microfilmer ses collections de journaux, il peut être préférable d'utiliser la caméra hybride pour effectuer en une seule opération la numérisation et le microfilmage. Toute manipulation de documents originaux étant lourde (en temps et en usure supplémentaire) il importe de ne les prélever et ne les utiliser qu'une seule fois.)

Il faut souligner que la qualité du microfilm est cruciale pour la qualité des images numérisées. Des outils de recherche améliorés dans les collections numériques exigent d'excellents résultats du traitement de reconnaissance optique de caractères qui ne peuvent être obtenus que grâce à la haute qualité technique du microfilmage. Autrement dit, la qualité du résultat final dépend de la qualité des opérations antérieures.

Quelques suggestions pour parvenir à une qualité améliorée

Les bibliothèques doivent se demander si un programme de microfilmage fondé sur des exigences de qualité doit faire partie de leur politique de numérisation et de conservation. Le niveau de la qualité doit être choisi d'après les besoins de la bibliothèque. Il est extrêmement important de respecter les normes prévues pour le microfilmage, et par exemple, d'utiliser les *Directives pour la réalisation de microfilms de conservation* publiées par IFLA. Il est également possible, si la bibliothèque le souhaite, de mettre en oeuvre dans un programme de microfilmage une numérisation largement automatisée avec un grand souci de qualité et la recherche en texte intégral.

Éléments déterminants du microfilm

Avoir un microfilm de contraste élevé est un élément déterminant, en particulier si l'on compare le résultat de la numérisation directe avec celle des microfilms, pour des textes gris sur papier jauni et cassant. Le texte se distingue alors clairement du fond. De faibles décolorations, de petites taches ou des pliures du papier sont éliminées, ce qui rend possible la recherche en texte intégral lorsqu'on utilise un logiciel de reconnaissance optique de caractères sur images bitonales. La résolution est tout à fait suffisante depuis le taux de réduction le plus bas jusqu'au taux de réduction 16 x, l'idéal étant un index de qualité de 18. Lorsque les taux de réduction dépassent 20 x, la reconnaissance optique de caractères risque de ne pouvoir se faire.

Toutefois, pour les documents photographiques, la numérisation à partir d'un microfilm ne sera jamais aussi bonne que la numérisation directe à échelle de gris. Il y a aussi une différence importante selon les diverses générations du microfilm utilisé : le duplicata négatif direct produit la numérisation la moins bonne pour les photos publiées dans les journaux.

Les résultats du projet nordique TIDEN

Au cours du projet TIDEN (voir ci-dessus) les résultats du traitement par logiciel OCR de journaux parus entre le début du 19e siècle et la fin du 20e siècle ont été comparés. Ces journaux furent numérisés à partir de microfilms 35 mm. Se sont dégagés quatre facteurs qui ont entraîné tout particulièrement le succès ou l'échec de la reconnaissance optique de caractères :

- la qualité du texte de l'original ; mises en pages multiples et qualité du papier
- le taux de réduction de la caméra
- le caractère et la dimension du caractère (les caractères romains donnent de bons résultats alors que le texte ancien en gothique exige de l'habileté et des efforts supplémentaires. Les caractères multiples comprenant des textes en gothique peuvent entraîner une interprétation inexacte du texte)
- la langue (le suédois donne de meilleurs résultats que le finnois). La reconnaissance optique de caractères de langues multiples mélangées dans une même zone ou dans une même phrase est plus difficile, voire parfois impossible à lire.

L'importance de la marque et de la résolution de la caméra, de la polarité et de la génération du film (première et deuxième générations) a été marginale. Toutefois il est à noter que les microfilms anciens faits dans les années 1950 et 1960 risquent de ne pas être de qualité suffisante pour la reconnaissance optique de caractères

alors que la qualité des films des années 1980 qui ont été réalisés selon les normes internationales est satisfaisante pour la reconnaissance optique.

Optimiser la numérisation à partir du microfilm

Il est possible d'obtenir une interrogation en texte intégral de textes numérisés à partir de microfilms aussi bonne qu'à partir d'une numérisation directe. Le potentiel du film n'est cependant pas toujours exploité au mieux en l'absence d'une bonne pratique de laboratoire lors de la préparation et de la prise de vue. C'est en programmant soigneusement chaque étape du processus qu'on améliore le niveau général.

La qualité des originaux est essentielle à la fois pour le microfilm et pour la numérisation. Aussi faudra-t-il rechercher les originaux dont l'état physique est le meilleur. Les documents très cassants doivent être manipulés sous la surveillance du personnel chargé de la conservation. Les journaux doivent être présentés en feuilles. Les volumes reliés doivent être microfilmés sur un plateau à bascule et sous presse-livre (dans ce cas la marge de fond doit être large de 1,5 cm à partir du fond du dos jusqu'au texte).

1. Les scanners de microfilms ont parfois des difficultés à identifier les différents formats des pages sur le microfilm. Cela est dû au fait que les scanners reconnaissent généralement une nouvelle vue à un endroit donné par son coin ou sa marge gauche. Dans ce cas l'identification peut être facilitée en marquant d'une bande le côté gauche de la vue sur la table photographique. Cette bande pourrait ensuite évoluer vers un code à barres comportant des informations lisibles en machine au sujet des images, telles que date, numéro (pagination) du journal, nouvelle édition, supplément, page répétée, etc. Une autre possibilité consiste à utiliser un logiciel pour résoudre les problèmes en question.
2. Lorsque le microfilmage est réalisé en vue de la numérisation, il est important que l'éclairage soit homogène à l'intérieur de chaque vue et tout au long du film ; il doit varier au maximum de 0.2 dans chaque vue et dans tout le film. Tout changement de densité à l'intérieur des vues entraînera une manipulation particulière de la part de l'opérateur numériseur, sinon la qualité de la reconnaissance optique de caractères sera réduite. Le changement de densité entre les vues nécessite un ajustement manuel des paramètres de numérisation pour chaque image, ce qui est plus coûteux et plus lent que la numérisation automatisée de film en continu.
3. Sur les microfilms anciens la distance entre les vues est parfois trop réduite pour que le scanner puisse séparer correctement les vues. Elle doit être suffisante et présenter la plus grande régularité.

En résumé, il faut rappeler que les normes nationales et internationales de microfilmage de conservation constituent une base pour la numérisation réussie de microfilms. La réalisation de films à partir de volumes déreliés et le taux de réduction du film sont les points essentiels à respecter dans l'effort pour obtenir la reconnaissance optique de caractères qui restitue l'original. Les variations de densité doivent être suffisamment réduites pour permettre d'utiliser la numérisation automatisée. Une bande blanche ou une barre doit indiquer le début de chaque image. Le pavé optique ou l'indexation lisible en machine par codes à barres pour la date, le numéro, la pagination du journal, ainsi que les images-témoins du microfilm, peuvent servir à améliorer l'automatisation de l'ensemble du processus depuis la numérisation jusqu'au montage des textes numérisés dans la base de données, etc. Des logiciels peuvent aussi conduire les procédés de numérisation et d'indexation.

Résultats économiques

La mise en oeuvre d'un programme de qualité pour la préparation des publications à microfilmer et pour le microfilmage proprement dit nécessitera à ses débuts un financement supplémentaire. En effet, l'organisation et la mise en place des deux parties du processus à la fois par la bibliothèque et par le prestataire de microfilms entraînent une hausse du coût. Mais elle a pour but de concevoir un programme de microfilmage pouvant servir de base aux futurs programmes de numérisation.

A titre d'exemples, des moyens financiers accrus pourraient être nécessaires pour abaisser les taux de réduction des films et pour le débrouillage des journaux reliés. Si l'on utilise des codes à barres, le remplacement des codes exigera une plus grande attention de la part de l'opérateur et risque ainsi de ralentir les cadences. La mise en place de nouveaux relevés de densité peut entraîner une plus grande exigence de qualité et décider du refilmage. La prise de vue de volumes reliés sur un plateau à bascule, sous presse-livre, pourrait nécessiter l'acquisition de nouveaux équipements. Toutefois, ces changements ont déjà été mis en place dans beaucoup de bibliothèques et ils peuvent être combinés avec les procédés habituels.

En dépit des dépenses initiales plus élevées, la haute qualité du microfilm avec les informations lisibles en machine permet de faciliter la numérisation et d'en réduire le coût. Les économies réalisées seraient éventuellement récurrentes car les progrès techniques et les exigences des usagers pourront entraîner la re-numérisation à une résolution plus élevée par profondeur de bit qu'il n'est possible de réaliser actuellement. Si le microfilm est d'excellente qualité, on n'aura pas besoin d'avoir recours à l'original qui, dans le pire des cas, n'existe plus. Donc, faites-le une seule fois mais faites-le bien !

La technologie viendra sans doute à bout partiellement, ou entièrement, de nos obstacles actuels. Concevoir un programme de microfilmage de qualité comme faisant partie de la stratégie de la bibliothèque facilitera les activités ultérieures de numérisation. Le microfilmage devrait faire partie intégrante du programme de numérisation de la bibliothèque.