ABBYY®

ETUDE DE CAS

ABBYY® Recognition Server

UNIVERSITY OF Southampton

L'université de Southampton améliore l'accès à sa collection en numérisant de volumineuses quantités de documents imprimés

Contexte

La bibliothèque de l'université de Southampton accueille environ 35 000 étudiants et membres de son personnel. Elle offre un accès à une très vaste collection de qualité supérieure : 1,5 millions de livres et plusieurs millions de pages de documents d'archive. La bibliothèque s'est récemment lancée dans un programme de numérisation d'un nombre important de textes clés via l'Unité de numérisation de bibliothèques (LDU, Library Digitisation Unit). La LDU est une opération phare dans le secteur de l'éducation et se spécialise dans la capture numérique d'un large éventail de documents pour les référentiels ou la distribution sur Internet via des liens du catalogue de la bibliothèque. La démarche de la bibliothèque universitaire consiste à offrir un accès complet aux documents numériques qu'elle crée, chaque fois que cela est possible et autorisé.

ABBYY FineReader, un produit logiciel conçu pour la numérisation et la numérisation ad-hoc, a été utilisé par la bibliothèque pendant un certain nombre d'années pour numériser une petite quantité de pages, et occasionnellement l'intégralité de livres. Ce programme est aussi utilisé pour la reconnaissance de texte et la reconnaissance optique de caractères (OCR, optical character recognition). La bibliothèque s'est rapidement rendue compte que pour automatiser le processus OCR, avec un rendement élevé, et atteindre son objectif, qui est de numériser un demi-million de pages par an, elle aurait besoin d'un produit plus robuste et capable de traiter automatiquement d'importants volumes de documents.

Solution

Pour trouver une solution permettant de traiter les documents imprimés, de les convertir en formats interrogeables, tels que le PDF et le PDF/A, et de numériser la création d'archives ou de dossiers, la bibliothèque a examiné diverses options disponibles sur le marché. Les critères suivants ont été examinés pour un certain nombre de produits: rapidité et qualité de l'OCR, gamme de formats et de compressions à l'exportation et possibilité d'intégrer des API/workflows. Après avoir examiné un certain nombre de produits, la bibliothèque a sélectionné ABBYY Recognition Server comme solution la plus performante.

La bibliothèque a choisi Recognition Server car ce logiciel répondait parfaitement à ses attentes: mise à disposition d'une reconnaissance de grande qualité pour les textes imprimés, un large éventail d'options d'exportation et une API ouverte permettant d'intégrer facilement d'autres programmes. La bibliothèque souhaitait intégrer Recognition Server au logiciel de workflow Goobi, développé par Intranda GmbH, de la LDU. Le logiciel de workflow de production Goobi est une application Internet qui gère et effectue un suivi des projets de numérisation de bibliothèques. Le niveau de qualité du service après-vente et une maintenance abordable ont été des atouts supplémentaires en faveur d'ABBYY.

La LDU utilise actuellement jusqu'à six scanners de livres et un scanner en ligne haut de gamme pour numériser les textes et les images faisant partie de sa collection. L'API Tickets XML d'ABBYY Recognition Server a été utilisée pour l'intégration à Goobi Workflow. Une fois les documents imprimés numérisés, Goobi gère la mise en attente de tâches Recognition Server puis contrôle les résultats. La production coordonnée de caractères peut être intégrée à une couche de présentation

À propos de l'université de Southampton

L'unité de numérisation de bibliothèques de l'université de Southampton propose un service de numérisation aux bibliothèques, aux archives et au secteur commercial. Depuis 2003, nous nous sommes spécialisés dans la capture numérique d'images et de texte issus de documents, que ces documents soient reliés ou non, afin de les conserver.

Contact

www.soton.ac.uk/library/ldu

ABBYY Europe GmbH
Elsenheimerstr. 49
80687 Munich, Allemagne
Tél: +49 89 51 11 59 0
Fax: +49 89 51 11 59 59
sales_eu@abbyy.com
www.ABBYY.com

Bureau France
4, rue Leroux
94100 Saint-Maur des Fossés
France
sales_france@abbyy.com
www.france.ABBYY.com

pour l'indexation et l'accès. Par conséquent, dès que l'utilisateur du scanner termine la numérisation d'un livre, les fichiers sont automatiquement déplacés vers chaque étape du workflow. Deux des plus importants programmes de numérisation du stock de la bibliothèque de l'université de Southampton comprenaient les projets financés par le Comité commun des systèmes d'information (JISC, Joint Information Systems Committee). Le JISC s'inspire des établissements d'enseignement supérieur et des universités du Royaume-Uni qui utilisent des technologies numériques de manière innovante et permet au Royaume-Uni de conserver sa position de leader mondial dans le secteur de l'éducation. Ces deux projets ont chacun généré plus d'un million d'images numériques:

Numérisation des documents papier du Parlement datant du 18ème siècle: ceci comprend les journaux de la chambre des Lords et de la chambre des Communes, les registres parlementaires, le Journal des sessions de la chambre des Lords, les factures et les lois d'intérêt privé et local de 1700 à 1834. Vous trouverez des exemples à l'adresse suivante :
http://www.southampton.ac.uk/library/ldu/parl18c.html

Numérisation des pamphlets datant du 19ème siècle: plus de 23 000 pamphlets datant du 19ème siècle provenant de bibliothèques de recherche couvrant le paysage économique et socio-politique de Grande-Bretagne ont été numérisés par la LDU et convertis au format PDF interrogeable. Les détails du projet et les entrées du catalogue sont disponibles à l'adresse suivante:
http://www.britishpamphlets.org.uk/

D'autres projets internes incluent:

Numérisation des thèses de doctorat: les copies numérisées de 20 000 de ces thèses récompensées de l'université de Southampton ont été converties en PDF interrogeable et sont à présent disponibles via le référentiel de recherche institutionnelle de l'université, Eprints Soton. Eprints Soton dispose de copies électroniques des résultats de recherche, notamment d'articles de journaux, de chapitres de livres, de documents de conférence et de thèses. Des manuscrits et des documents papier non publiés sont également disponibles. L'intégralité du texte de nombreux de ces éléments est gratuitement disponible en fonction des autorisations liées au droit d'auteur et des autorisations accordées à l'utilisateur final. **http://eprints.soton.ac.uk/**

Numérisation des documents parlementaires relatifs à l'Irlande (EPPI): environ 15 000 documents parlementaires, de 1801 à 1922, ont été numérisés et mis à disposition sous la forme de fichiers PDF interrogeables via des liens du catalogue de l'université.

Numérisation de textes de cours: la bibliothèque convertit les documents imprimés au format PDF interrogeable afin de soutenir les cours nécessitant une grande quantité de textes. Dans le cadre de la licence accordée par l'organisme de concession de licences de droits d'auteur (CLA, Copyright Licencing Agency) du Royaume-Uni, l'accès aux fichiers PDF interrogeables se limite aux utilisateurs de l'université via le catalogue de la bibliothèque.

La collection Richard Rutt: une collection de livres de tricot datant du 19ème siècle provenant de la bibliothèque de références artistiques de l'école Winchester a été numérisée et mise à disposition sous la forme de fichiers PDF interrogeables via Internet.

Conclusion

La reconnaissance optique de caractères puissante et précise est une partie intégrante des activités de l'unité de numérisation de bibliothèques. Grâce à l'augmentation de la production engendrée par ABBYY Recognition Server, le personnel de la bibliothèque est libéré du travail ennuyeux que constitue l'OCR manuel. Des millions de pages de documents provenant de la collection de l'université de Southampton sont disponibles en ligne dans des formats numériques pour les étudiants et pour le reste du monde.

La réussite de la numérisation peut également être attribuée à la souplesse d´administration du produit et à la facilité de son intégration aux processus de bibliothèques existants. « La possibilité d'intégrer ABBYY Recognition Server à notre workflow était capitale, » a déclaré Julian Ball, administrateur de section de l'unité de numérisation de bibliothèques.

« L'installation d'ABBYY Recognition Server a été rapide. Les premières réponses d'ABBYY et de leurs équipes de support vis-à-vis de toutes les questions posées ont été rapides, avec un suivi de bonne qualité. Nous avons été très satisfaits des résultats et nous avons hâte d'utiliser le produit pour continuer à remplir nos objectifs en termes de numérisation. »

A propos d'ABBYY

ABBYY est un leader en matière de développement de technologies de reconnaissance de documents, de conversion de documents, d'acquisition de données et de linguistique.

Les produits ABBYY sont composés de: ***FineReader*** *et* ***PDF Transformer*** *– logiciels destinés à l'utilisateur final pour la conversion de documents* ***Recognition Server:*** *une solution OCR et de conversion PDF sur serveur;* ***FlexiCapture:*** *programmes d'acquisition de données pour le traitement des formulaires, des documents semi-structurés et non structurés; Les* ***SDK*** *de* ***FineReader Engine*** *qui couvrent tout le spectre des technologies de reconnaissance d'ABBYY; et* ***Lingvo:*** *une gamme de dictionnaires logiciels.*

Pour plus d'information:
www.abbyy.com

ABBYY®

ABBYY Europe GmbH
Elsenheimerstr. 49
80687 Munich, Allemagne
Tél: +49 89 51 11 59 0
Fax: +49 89 51 11 59 59
sales_eu@abbyy.com
www.ABBYY.com

Bureau France
4, rue Leroux
94100 Saint-Maur des Fossés
France
sales_france@abbyy.com
www.france.ABBYY.com