ABBYY®

FALLSTUDIE

ABBYY® FineReader Engine

invent

Das Deutsche Patent- und Markenamt (DPMA) verwendet ABBYY FineReader Engine zur Erschließung der Patendokumentation als Volltext

Von der Blättersekunde zur Lesemaschine

Anfang der 90ziger Jahre begann das Deutsche Patent- und Markenamt DPMA mit dem Start des Projekts DEPATIS, mit dem die Patentprüfung von reiner „Papierverarbeitung“ auf modernste Technologie umgestellt wurde. Damals war man zufrieden damit, den Prüfern 20 Millionen Patentdokumente als Faksimile am Bildschirm zur Verfügung stellen zu können – auch das bei der verfügbaren Technologie keine leichte Aufgabe. Die „Blättersekunde“ war der Schlüsselbegriff, die Prüfer sollen in weniger als einer Sekunde eine beliebige Seite eines beliebigen Dokuments aus dem riesigen Bestand auf dem Bildschirm haben. Es entstand mit dem System DEPATIS eines der modernsten Recherche- und Archivsysteme überhaupt. Damit können die Patentprüfer des DPMA und – über das Internet – auch die Öffentlichkeit in einem der weltweit umfassendsten Bestände an Patentdokumenten recherchieren.

Als der Ausbau des Systems DEPATIS zur so genannten Vollausstattung geplant wurde, lebte der Gedanke der Volltexterschließung wieder auf. Technischer Fortschritt und das gesammelte Wissen über die Patentdokumente ließen ein solches Vorhaben machbar erscheinen. Ein kleinerer Bestand an Volltextdaten war bereits vorhanden: die Deutschen Patentschriften wurden bereits seit 1987 als sog. DATIMTEX-Dokumente textcodiert erfasst. Insgesamt waren etwas über 1 Million Dokumente vorhanden, in denen die Prüfer im Volltext recherchieren konnten. Die Problematik einer Patentrecherche liegt aber im Anspruch auf Vollständigkeit. Es müssen alle Stellen in allen verfügbaren Dokumenten gefunden werden, die den zu prüfenden Sachverhalt beschreiben oder mit ihm in engem Zusammenhang stehen. Und das bedeutet, dass für eine sinnvolle Recherche der Dokumentenbestand möglichst vollständig textcodiert zur Verfügung stehen sollte. Es galt also, alle Dokumente im Volltext zu erfassen. Von den inzwischen 30 Millionen im DEPATIS-Archiv vorhandenen Faksimiledokumenten lagen neben über einer Million deutscher Volltexte noch knapp 7 Millionen Zusammenfassungen von japanischen Dokumenten textcodiert vor, dennoch handelte es sich angesichts eines *„Restbestandes“* von mehr als 20 Millionen Dokumenten um eine enorme Aufgabe.

Umfangreiche Planungen und eine intensive Sondierung möglicher Datenlieferanten musste das DPMA durchführen, bis schließlich im September 2002 das Projekt mit dem einfachen Namen »Vollausstattung – Los 3« gestartet wurde. Nach einer europaweiten Ausschreibung wurde HP auf Grund des wirtschaftlichsten Angebotes vom DPMA mit dem Ausbau des DEPATIS Recherchesystems beauftragt. Schwerpunkt des Projektes war die Erweiterung des im Volltext recherchierbaren Datenbestands von 1,2 Millionen auf rund 12 Millionen Dokumente. Dabei hatte das Amt die Anforderung, dass trotz Vervielfachung der Menge an Volltexten gleich bleibende Antwortzeiten für die Recherche und sogar bessere Antwortzeiten bei den Pflege-Operationen (Prüfstoffpflege) erzielt werden. Die Aufgabe wurde von HP in mehreren Schritten gemeistert:

Über HP Services – Consulting & Integration

HP Services – Consulting & Integration unterstützt Unternehmen mit einem umfassenden Leistungsrange bei der Planung und Realisierung innovativer, offener IT-Lösungen und aktivem Change Management. Von der Beratung über die Systemintegration bis zur Projektimplementierung steht Consulting & Integration für messbare Erfolge. Mit einer globalen Präsenz von über 4800 Consultants, davon 400 allein in Deutschland, steht Consulting & Integration für technische Führerschaft, umfassendes Wissen und professionelles Projektmanagement. Weltweit konsistentes Wissensmanagement garantiert dem Kunden Effizienz und aktuellstes Know-how durch globalen Erfahrungsaustausch zwischen HP Consultants. Consulting & Integration kooperiert mit den Besten der IT-Branche, um für jede Herausforderung individuell optimale Lösungen zu garantieren.

Kontakt

Hewlett-Packard GmbH
HP Services – Consulting & Integration
Herrenberger Straße 140
71034 Böblingen
Tel.: 07031-140
Fax: 07031-142999

ABBYY Europe GmbH
Elsenheimerstr. 49, D-80687 München
Tel: +49 89 51 11 59 - 0, Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com

ABBYY® FineReader Engine

Speicherung der Daten im XML-Format

Zum ersten wurde für die Speicherung der textcodierten Dokumente XML als einheitliches, zukunftsweisendes Datenformat festgelegt, und die entsprechenden Umstellungen im Archivsystem vorgenommen. Insbesondere mussten die vorhandenen Volltext-Daten in das neue Format konvertiert werden, und es mussten die Datenimport und -exportverfahren angepasst werden, um die laufende Übernahme der neuen Dokumente im Volltext sicher zu stellen. Im Rahmen dieser Umstellungen wurde auch die Internet-Schnittstelle für den Download der Daten durch die Öffentlichkeit um eine XML-Schnittstelle erweitert.

Erkennung der Dokumente mit der OCR Software ABBYY FineReader Engine

Danach wurde nach der passenden OCR Software zur Erkennung der Dokumente gesucht. Ulrich Merz, Projektmanager von HP Services – Consulting & Integration, sagt über die Auswahl der OCR Engine: *„In einer mehrwöchigen Evaluierungsphase wurden verschiedene Desktop-Produkte geprüft. Entscheidungskriterien waren die Erkennungsgenauigkeit, die Erkennungsgeschwindigkeit, die Sprachunterstützung und vor allem das Vorhandensein einer Programmierschnittstelle zur Einbindung in ein automatisiertes Verfahren. Geprüft wurden die Produkte gegen eine Auswahl von repräsentativen Patentdokumenten der vergangenen 100 Jahre. In diesen Tests hat uns FineReader Engine von ABBYY am meisten überzeugt.“*

Auf der Basis des Produkts FineReader der Firma ABBYY, kombiniert mit einer automatischen, intelligenten Nachbearbeitung der Ergebnisse wurde eine Produktion von Volltexten aus den vorhandenen Faxdokumenten aufgebaut. Das Verfahren weist eine sehr hohe Ergebnisqualität auf, so gibt es z.B. die automatische Erkennung und geeignete XML-Präsentation von Mehrspaltigkeit, Tabellen, bestimmten Textabschnitten („Subdokumente“) u. a. mehr. Nicht relevante Textbestandteile wie z.B. Kopf-/Fußzeilen, Aufdrucke/Stempel und Ähnliches werden ausgesondert.

Das OCR-Verfahren ist in einer modernen Mehrschichtarchitektur implementiert. Zum Einsatz kommen zwei redundant ausgelegte HP Alphaserver DS20 mit BEA-Weblogic-Applicationservern und derzeit 7 HP Proliant Server für die eigentliche OCR-Wandlung. Dadurch wird ein Durchsatz von 4.500 Seiten pro Stunde bzw. bis zu 40 Millionen Seiten pro Jahr erreicht. Die gewählte Architektur ist einfach durch weitere OCR-Server skalierbar.

Das OCR-Verfahren realisiert – neben der eigentlichen Texterkennung – eine automatische, intelligente Nachbearbeitung der Daten und erzeugt so die strukturierten, XML-codierten Textdaten.

Bis Juni 2005 wird die Umwandlung von ca. 3 Millionen deutschen Patentdokumenten abgeschlossen sein. Anschließend werden weitere Dokumente weiterer Länder gewandelt. Der Volltextbestand im System DEPATIS wächst rasant. Das „Einfüttern“ der Daten erfolgt ohne Unterbrechungen des DEPATIS-Betriebs, die durch OCR erzeugten Daten stehen sofort den Anwendern zur Verfügung. Es gilt, über 20 Millionen Dokumente aus verschiedensten Ländern als Volltext in das System einzubringen. Das OCR-Verfahren wird für alle Bereiche angewendet, in denen keine Volltextdaten aus anderen Quellen in entsprechender Qualität verfügbar sind. Ausgenommen werden nur wenige Länder, z.B. Länder, bei denen eine Volltextrecherche nicht zur Anwendung käme, weil kein Rechercheur die Sprache ausreichend beherrscht. Von den 4,2 Millionen Deutschen Patentdokumenten lagen zu Beginn der Umwandlung bereits 1,2 Millionen als Volltext vor. Die Komplettierung des Volltextbestandes der deutschen Dokumente wird bis Juni 2005 abgeschlossen sein.

Über ABBYY

ABBYY ist ein führendes Unternehmen in der Entwicklung von Technologien für Dokumenterkennung, Dokumentumwandlung, Data Capture und Linguistik. Zum Produktportfolio von ABBYY gehören: FineReader OCR und PDF Transformer – Endanwender-programme zur Umwandlung von Dokumenten; Recognition Server – eine serverbasierte Lösung für OCR und PDF-Umwandlung; FlexiCapture und FormReader – Data Capture Programme zur Verarbeitung von Formularen, semi- und unstrukturierten Dokumenten; FineReader Engine SDKs mit dem gesamten Leistungsumfang der ABBYY OCR-Technologien; Lingvo – eine Serie von elektronischen Wörterbüchern.

Mehr Informationen über ABBYY unter www.ABBYY.com

ABBYY®
ABBYY Europe GmbH
Elsenheimerstr. 49, D-80687 München
Tel: +49 89 51 11 59 - 0, Fax: +49 89 51 11 59 - 59
sales_eu@abbyy.com
www.ABBYY.com