

AxiCom-PR, Nr. AB 16/04, Dezember 2004

Das Projekt Gutenberg-DE verwendet ABBYY FineReader XIX zur Erkennung antiquarischer Bücher

ABBYY FineReader XIX digitalisiert Bücher in Frakturschrift

München, 07. Dezember 2004 – ABBYY, führender Anbieter von Technologien zur Dokumenten-Erkennung und linguistische Technologien, und das Projekt Gutenberg-DE starten heute in Zusammenarbeit mit der Berlin-Brandenburgischen Akademie der Wissenschaften ein gemeinsames Projekt. Unter dem Arbeitstitel „GaGa – Gemeinsam an Gutenberg arbeiten“ werden Frakturtexte sowie OCR-Rohdaten über das Internet zum verteilten Korrekturlesen angeboten. Seit Gründung des Projektes Gutenberg-DE im Jahr 1994 sind mehrere zehntausend unbezahlte Arbeitsstunden erbracht worden, um Literatur kostenlos ins Internet zu stellen. Ab sofort können durch Einsatz der OCR Software ABBYY FineReader XIX jetzt auch Bücher, die in Frakturschrift gedruckt wurden, erkannt und online gestellt werden.

Internet-Community digitalisiert antiquarische Literatur

Unter der Internetadresse www.gaga.net hat das Projekt Gutenberg-DE das verteilte Korrekturlesen von Büchern gestartet. Die Idee ist simpel und verspricht großen Erfolg: Jeder Teilnehmer am Projekt sieht das Bild einer Buchseite und den von einem OCR-Programm erkannten Text dieser Seite, der noch einige Fehler enthalten kann. In etwa drei Minuten ist solch eine Seite korrigiert. Diese Seite wird an die Textdatenbank zurückgegeben und sofort erhält man die nächste Seite zur Korrektur. Das Lesen oder Laden eines kompletten Buchtextes ist im Projekt GaGa übrigens nicht möglich, weil die Folgeseite schon von einem anderen korrigiert und damit solange gesperrt wird, bis die Korrektur beendet ist. Auf eine fertig korrigierte Seite kann ebenfalls nicht mehr zugegriffen werden.

Das Projekt Gutenberg-DE hat täglich 30.000 Besucher, die auf der Suche nach Texten deutscher Klassiker sind. „Wenn nur jeder hundertste Leser auch nur eine Seite korrigiert, können wir jeden Tag ein fehlerfreies Buch von 300 Seiten fertig stellen“, meint der Projektleiter Gunter Hille, der vor 10 Jahren das Projekt Gutenberg-DE begann. Und diese Aussage ist noch bescheiden, denn das US-Pendant (www.pgdp.net) hat bereits eine Seitenleistung von knapp 6.000 Seiten pro Tag erreicht.

„Wir können die Digitalisierung alter Bücher revolutionieren, die noch nicht in digitaler Form vorliegen, denn bisher scheuen Verlage die hohen

Erstellungskosten, selbst wenn die Datenerfassung in Billiglohnländern erfolgte“, so der Projektleiter.

Gutenberg-DE

Das Projekt Gutenberg-DE wurde 1994 als Freizeitprojekt begonnen, als es nur wenige deutschsprachige Texte im Internet gab. Bis heute sind mehrere zehntausend Arbeitsstunden für das Projekt vom Gutenberg-Team aufgebracht worden. Seit über zehn Jahren stellt das Projekt Gutenberg-DE kostenlos für Jeden Literatur ins Internet. Das Team hat in dieser Zeit unter Mithilfe von freiwilligen, unbezahlten Helfern die größte deutschsprachige Online-Literatursammlung mit bis zu 3,3 Millionen Seitenabrufen im Monat aufgebaut. Bisher wurden ca. 420.000 Textseiten digitalisiert, darunter 14.000 Gedichte und 1.700 vollständige Romane, Erzählungen, Novellen. Weitere Informationen finden Sie im Internet unter <http://gutenberg.spiegel.de/index.htm>

Berlin-Brandenburgische Akademie der Wissenschaften (BBAW)

Das an der BBAW beheimatete „Digitale Wörterbuch“ (DWDS) hat in der ersten Projektphase ein Textcorpus der deutschen Sprache des 20. Jh. im Umfang von über einer Milliarde Textwörtern (in über 2 Millionen XML-Dokumenten). Das Textcorpus ist mittels einer linguistischen Suchmaschine abfragbar. Die Bedeutung für die sprachwissenschaftliche Forschung wird durch mehr als 1 Million Seitenaufrufe (seit September 2004) dokumentiert. Die Textbasis soll auf der Basis der Quellenbibliothek des Deutschen Wörterbuchs von Jacob und Wilhelm Grimm sukzessive ausgebaut werden. Weitere Informationen finden Sie unter www.dwds.de

ABBYY Europe GmbH

ABBYY Europe ist ein Mitglied der ABBYY Gruppe und veröffentlicht und vertreibt ABBYY-Produkte in Westeuropa. ABBYY (ABBYY Software House) entwickelt linguistische Software und Software für Künstliche Intelligenz, darunter eine komplette Produktlinie für OCR und Texterkennungsprogramme und Lösungen. Das Produkt-Portfolio von ABBYY bietet: FineReader OCR Systeme – eine Produktfamilie für Endkunden und SDKs zur Erkennung von gedruckten Texten, Tabellen und Formularen; FormReader – ein ICR Programm zur Erkennung und Verarbeitung von Formularen, die per Hand ausgefüllt wurden; ABBYY Retrieval & Morphology Engine – Tools für Entwickler zur Integration von Volltextsuche und linguistischen Funktionen in externe Applikationen. Zu den Lizenznehmern der Erkennungs-Technologien von ABBYY zählen Siemens Nixdorf, Samsung Electronics, C-Technologies, Sumitomo Electric Systems, Banctec, BenQ, Documentum, EPSON, Freedom Scientific, Hewlett-Packard, Kurzweil, Kofax, Lexmark, Microtek, NewSoft, Notable Solutions, Panasonic, Stellent, Toshiba, Verity, Neurascript, Saperion und SWT. Die Zentrale der ABBYY Gruppe befindet sich in Moskau, Büros gibt es in der Ukraine (ABBYY Ukraine), den USA (ABBYY USA, Fremont, CA), Großbritannien (ABBYY UK, Bishops Stortford, England) und Deutschland (ABBYY Europe, München). Weitere Informationen über ABBYY finden sich im Web unter www.abbyy.com oder direkt bei ABBYY Europe GmbH, Anglerstraße 6, 80339 München, E-Mail: sales@abbyeu.com.

Weitere Informationen:

Hille & Partner

abc.de Internet-Dienste

Gunter Hille

Spaldingstr. 160 D

20097 Hamburg

Tel: 040 - 899 75 55

Fax: 040 - 899 75 56

E-Mail: hille@abc.de

ABBYY Europe GmbH

Gino Boscia

Anglerstraße 6

D - 80339 München

Tel.: 089/511 159-17

Fax: 089/511 159-59

E-Mail: press@abbyeu.com

Pressekontakt:

AxiCom GmbH

Marcus Birke

Junkersstraße 1

D - 82178 Puchheim

Tel.: 089/800908-26

Fax: 089/800908-10

E-Mail: marcus.birke@axicom.de

Web: www.axicom.de