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

The Gutenberg-DE project applies ABBYY FineReader XIX to recognise ancient books

Gothic printed books are now digital thanks to ABBYY FineReader XIX

Munich, 07. December 2004-ABBYY, a leading provider of technologies for document recognition and linguistic technologies, and the Gutenberg-DE project are today starting a conjoint project in cooperation with the Berlin-Brandenburg Academy of Science. Under the working title "GaGa –working together at Gutenberg" Gothic printed texts as well as OCR raw data are being offered for distributed proofreading on the Internet. Since the launch of the Gutenberg –DE project in 1994 lots of non-paid overtime hours have been invested, so that literature can be posted for free on the Internet. From now on books printed in Gothic font can be recognised and uploaded to the Internet using ABBYY's OCR software FineReader XIX.

Internet community digitalizes ancient literature

www.gaga.net is the internet address that the Gutenberg-DE project set up to enable the distributed proofreading of books. The idea is simple and could be a huge success: each user can see one page as an image of the original book page and the text form of the page which has been recognised by the OCR Software. This page can still have some mistakes, which can then be corrected by the user. It takes about 3 minutes for one page to be recognised. That page is then sent back to the database and you immediately receive the next page to be corrected. To read or download the whole text of a book is impossible within the scope of the GaGa project as the next page could be being read and corrected by another user and during that time it is blocked until the page is corrected. It is also not possible to access and correct an already finalised and approved page.

The Gutenberg-DE project homepage is visited by around 30.000 visitors searching for texts from German classics. "If just 1 out of every 100 visitors corrects one page, we could end up with 300 page book each day without any mistakes", says Gunther Hille, the project manager who started the Gutenberg-DE project 10 years ago. And this is actually a modest conclusion if you take into consideration that the US equivalent (www.pgdp.net) has already managed to achieve an average of almost 6.000 pages per day.

"This could revolutionize the digitalisation of ancient books (which are currently not available in digital format), as publishers have so far avoided doing this because of the high costs involved even when the data capture is done in low wage countries" says the project manager.

Gutenberg- DE

The Gutenberg-DE project was started as a part time project, when there were just a few German texts on the Internet. Currently some ten of thousands of working hours have being invested into this project. For over ten years the Gutenberg-DE project has been putting free literature up for everybody on the Internet. The team, with the help of unpaid volunteers, has in that time succeeded in putting together the biggest

German collection of online literature with about 3,3 millions site visits per month. About 420.000 pages have been digitalised, amongst them poems and 1.700 complete novels, tales, short novels. For more information please see http://gutenberg.spiegel.de/index.htm

Berlin-Branderburg Academy of Science (BBAW)
The BBAW "Digital Dictionary" (DWDS) had, in its first phase of the project, a body of German texts from the 20th century containing 20 billion words (in more than 2 million XML-documents). The content of these texts is searchable with a linguistic search machine. The importance for philological research is being proved by the fact that it has had over 1 million site visits (since September 2004). The text basis is to be continuously completed on the basis of the source library of the Jacob and Wilhelm Grimm brothers German Dictionary. For more information please see www.dwds.de

ABBYY Europe GmbH
ABBYY Europe GmbH is a member of the ABBYY Group, and distributes ABBYY products in Western Europe. ABBYY (ABBYY Software House) develops linguistic and Artificial Intelligence (AI) software providing a full line of OCR and text-processing programs and solutions. ABBYY's products include: FineReader OCR systems – a family of end user programs and development tools for recognition of printed texts, tables and forms; FormReader – an ICR program for recognition and processing of hand-printed forms; and ABBYY Retrieval & Morphology Engine – a set of developer's tools for integration of full-text search and linguistic capabilities into external applications. Companies that license ABBYY recognition technologies include Siemens Nixdorf, Samsung Electronics, C-Technologies, Sumitomo Electric Systems, Banctec, BenQ, Documentum, EPSON, Freedom Scientific, Hewlett-Packard, Kurzweil, Kofax, Microtek, NewSoft, Notable Solutions, Panasonic, Stellent, Toshiba, Verity, Neurascript, Saperion, and SWT. ABBYY Group headquarters are located in Moscow with offices in Ukraine (ABBYY Ukraine), the USA (ABBYY USA, Fremont, CA), the UK (ABBYY UK, Bishops Stortford, England), and Germany (ABBYY Europe, Munich). ABBYY Europe is located at Anglerstrasse 6, 80339 Munich, Germany and can be reached at 49 (0) 89 51 11 59 -0 or via email at sales@abbyyeu.com. For more information about ABBYY, please visit the company's website at www.abbyy.com.