



FineReader 8.0 Professional Edition

FormReader Desktop Edition 6.5

### ABBYY FINEREADER 8.0 PROFESSIONAL

OCR software podporující i mimoevropské jazyky

**Minimální požadavky** – CPU 500 MHz, 128 MB RAM, 350 MB na HD, Windows 2000/XP/2003

**Cena** – 5236 Kč vč. DPH  
(Corporate Edition 10 472 Kč)

### ABBYY FORMREADER 6.5 DESKTOP EDITION

OCR software pro hromadné načítání vyplněných formulářů

**Minimální požadavky** – CPU 350 MHz, 128 MB RAM, 120 MB na HD, Windows 98 až 2003

**Ceny (bez DPH)** – licence na zpracování až 25 000 str. A4 během 1 roku = 19 200 Kč, po roce nutno obnovit (16 000); licence na zpracování až 10 000 str. A4 měsíčně bez časového omezení = 38 400 Kč.

**Vyrábí** – ABBYY Software House, Moskva, Rusko

**Poskytl** – NUPSESO CZ, [www.nupseso.cz](http://www.nupseso.cz)

# Aby stroje lépe četly

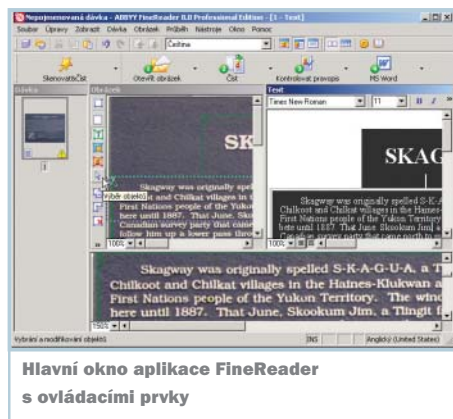


Z produktů ruské firmy ABBYY byl v Chipu recenzován program FormReader 4.1 a program FineReader 7. Nyní jsme otestovali nejnovější verze obou těchto programů a posoudili jsme novinky, které přinášejí.

**Text:** Miroslav Herold

## FineReader 8.0

Při instalaci programu FineReader (dále jen FR) se doporučuje instalovat pouze ty jazyky



Hlavní okno aplikace FineReader s ovládacími prvky

rozpoznávání, které budou skutečně využívaty, protože velikost prostoru na disku závisí hlavně na jejich počtu. Při šesti jazycích zabírala instalace 112 MB diskového prostoru. Soubory podporující rozpoznávání určitého jazyka jsou různě náročné: esperanto 1 kB, angličtina včetně rozšiřujících slovníků (právo + medicína) 2,04 MB, čeština 2,4 MB, ruština 4 MB apod.

## Přehled novinek proti verzi 7.0

- Přesnost rozpoznávání dokumentů nízké kvality (faxy, skenování < 200 dpi) byla zlepšena – podle údajů výrobce až o 30 %.
- FR nyní aplikuje duální přístup k PDF dokumentům. Určí, zda je text v dokumentu vsazen (pak jej pouze extrahuje), pokud není, nasazuje algoritmy OCR. Toto zjišťování se provádí pro každý blok.
- Pokud v PDF souboru existují vnitřní propojení nebo hypertextové odkazy, FR je v cílovém dokumentu MS Word obnoví.
- Jestliže chceme konvertovat pouze některé stránky z dokumentu, lze je nyní vybrat už při načítání souboru.

■ U obrázků z digitálního fotoaparátu byl testován špatně čitelný snímek, a i ten se po několika opravách podařilo převést do DOC. Když se provedla úprava perspektivního zkreslení, přesnost čtení se podstatně zlepšila.

■ Funkce vyrovnání řádků přináší zlepšené rozpoznání při zpracování vyfocených textů i při skenování z knihy (odstraní zahnutí řádků u hřbetu).

■ Naskenované dokumenty lze ukládat do formátu LIT, kompatibilního s aplikací MS Reader. Testován byl následný přenos do PDA, naskenovaný dokument byl dobře čitelný i na malé obrazovce.

■ Slovníky méně obvyklých slov, postupně vytvářené při psaní specifických textů v MS Wordu, lze nyní přiřadit k procesu rozpoznávání.

■ FR podporuje celkem 179 jazyků, podporu korektora má 36 jazyků (přidána slovenština a baškirština). Mezi jazyky bez korektora je celá řada takových, které používá méně než 10 000 lidí.

■ Bonusová aplikace ABBYY Screenshot Reader. Tu si jistě oblíbí všichni, kdo častě-



Fotografie s obtížně umístěným textem, který však byl i přesto ještě rozpoznán.



Upravená fotografie z předchozího obrázku – rozpoznání proběhlo bez problémů.

➔ Ji pracují s různorodými zdroji a musejí je slučovat do jednoho výsledného dokumentu. Umožňuje totiž přenášet do schránky nebo do Wordu text či obrázky z formátů, které buď přenos vůbec nedovolují, nebo kde je značně složitý. Text by se mohl načíst přes FR (otevřít soubor, definovat blok, přečíst), avšak přes Screenshot Reader je to mnohem rychlejší. Stačí pouhý výběr textu a ten se buď očitve ve schránce, nebo je převeden do DOC (předdefinováno na začátku práce). Podobné je to po excerptci z internetu – tam také některé stránky nepodporují kopírování.

**Praktické poznatky a připomínky**

Testování bylo zaměřeno spíše na problémy při úpravách načteného dokumentu a při jeho exportu. Se samotným rozpoznáváním textu jsou – díky dobře zpracovaným algoritmům – problémy opravdu jen zřídka. Dokonce ani barevný tisk na podkladě jiné barvy, tradiční slabina OCR aplikací, nedělal potíže.

Problém působí načítání textů nepravidelně rozmístěných na stránce okolo obrázku (častý případ v technických manuálech). FR některé texty interpretuje jako slopcový text s proměnnou výškou řádky, jiné jako rámy (frames). Pro následné zpracování by bylo výhodnější soustavné používání textových polí (text box). Rámy totiž neumožňují například transparentní podklad textu, což vadí u textových popisek umístěných na jiném než na bílém podkladu.

**Zhodnocení FineReaderu**

FineReader 8.0 už není jen OCR aplikací, nabídka jeho funkcí pokrývá široké pole práci s texty. Pro českého uživatele je zajímavá i bohatá paleta korektorů pro veškeré jazyky států obklopujících ČR i pro mnoho dalších.

Počet chybných překladů v uživatelském rozhraní programu, vytýkaný u verze 7.0, se podstatně snížil, některé však zůstaly.

Manuál má 112 stran, je v češtině a dostatečně podrobně popisuje práci s aplikací. Uživateli je má k dispozici buď v tištěné verzi, nebo ve formátu PDF.

**FormReader 6.5**

Na rozdíl od programu FineReader, který poslouží i jednotlivým uživatelům, je program FormReader určen pro sféru velkých podniků, kde se zpracovávají data z formulářů ve velkém (a spíše ještě pro oblast státní správy) – je připraven pro zpracování více než tisíce formulářů v jedné pracovní směně. Typickým příkladem je vyhodnocování zkušebních testů ve školství anebo různých dotazníků ve státní správě, dotazníků o průzkumu trhu, načítání dat z provozních záznamů v průmyslových podnicích či dat z předtištěných smluv pro pojišťovny.

**Novinky verze 6.5**

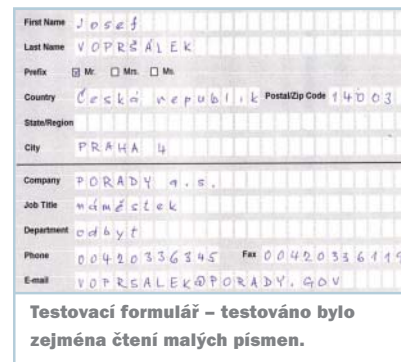
- Podpora vícestránkových dokumentů o pevném počtu stran a daném pořadí.
- Otevírání obrazů ve formátu PDF a ukládání do tohoto formátu.
- Skenování používající ovladače ISIS, což dovoluje použít více než 250 druhů vysokorychlostních skenerů.
- Zlepšené možnosti ověření a ověřovacích skupin – ke každému ověření je nyní možné přiřadit libovolně abecedně-číselcové hodnoty; ověření lze seskupit do skupin bez ohledu na umístění polí v rámci obrazu.
- Přidány slovníky pro litevštinu.
- Ověřovací pravidla je nyní možné vytvářet v jazyce Visual Basic.
- Rozpoznávání ručně psaných znaků (ICR) zvládá 22 jazyků s podporou slovníku (včetně češtiny a slovenštiny) a dalších 69 jazyků s latinkou, azbukou nebo řeckým písmem.

- Rozpoznávání tištěného písma (OCR) zvládá 172 jazyků, z toho 34 s podporou slovníku.
- Rozlišení čárových kódů EAN 8, EAN 12 a dalších dvanácti kódů.

**Zhodnocení FormReaderu**

K testování byl použit dotazník, obsahující písmenová a číselná pole. Pokud lze zadat vyplnění jen velkými písmeny, zvýší se správnost rozpoznání – u malých tiskacích písmen se někdy zaměňují znaky c/e a m/n. I u nekvalitně vyplněných formulářů je však vyhodnocení přesné a hlavně velmi rychlé – podle údajů firmy ABBYY až dvacetkrát rychlejší než zvládne nejzkušenější operátor „ručně“.

Při pořizování produktu je nutno uvážit, že jeho pořizovací cena je jen částí celkových nákladů. Stojí proti sobě náklady na manuální zpracování, zahrnující značnou chybovost a dlouhou dobu zpracování, proti nákladům na pořízení softwaru a mzdovým nákladům pracovníků podílejících se na projektu. Podstatným přínosem je kratší doba potřebná ke zpracování. Určité typy projektů lze realizovat už jen výhradně s nasazením takového aplikace – například sčítání obyvatelstva/bytů apod. ■ ■ ■



Testovací formulář – testováno bylo zejména čtení malých písmen.