

# PDF KONVERZIJA

## Kako se OCR engine može iskoristiti za neke druge stvari

**V**erujemo da svi znate za poznati OCR sistem koji se odaziva na ime FineReader. Kompanija ABBYY koja je proizvođač ovog softvera u svojoj paleti proizvoda ima još nekoliko korisnih alata koji možda na prvi pogled nisu povezani sa FineReaderom, ali u osnovi sadrže istu tehnologiju. Jedan od takvih alata je i PDF Transformer. Pre nego što krenemo sa opisom samog programa, moramo da se osvrnemo na to zašto je uopšte odlučeno da se napravi jedan ovakav alat. PDF je danas standard koji ne zavisi od platforme i neka istraživanja govore da se gotovo 10% dokumenata na Internetu može pronaći upravo u ovom formatu. Ako tome dodamo da je ovaj format osnova za čuvanje dokumenata u mnogim kompanijama, jasno je da postojanje jednog sveobuhvatnog alata za rad sa PDF-ovima neophodno za efikasno funkcionisanje u takvim uslovima.

ABBYY je uvideo da pregled i štampanje PDF dokumenata u poslovnom okruženju nije uvek dovoljno. Često je potrebno da se informacije koje se nalaze u PDF dokumentima ponovo iskoriste, ili da se neki dokumenti pretvore u PDF format. I dok za ovo drugo postoji mnogo rešenja, "čupanje" podataka iz PDF-ova nije ni malo jednostavno, naročito ukoliko je potrebno sačuvati izgled stranice. Ponekad autori PDF dokumenata ne dozvoljavaju čak ni selektovanje, kopiranje ili štampanje podataka iz dokumenata, pa se pred korisnike takvih dokumenata postavljaju gotovo nemogući zadaci.

Možda je tako bilo do pojave PDF Transformer-a. Ovo je jedan od retkih kompletnih alata koji u sebi objedinjavaju izradu PDF-ova, ali i izvlačenje podataka iz istih ili, možda je pravilnije reći konverziju PDF dokumenata u neki drugi format koji je jednostavniji i lakši za dalju obradu. Zato se PDF Transformer sastoji iz dva nezavisna alata, od kojih jedan kreira PDF dokument, a drugi konvertuje PDF-ove u neki drugi tekstualni format.

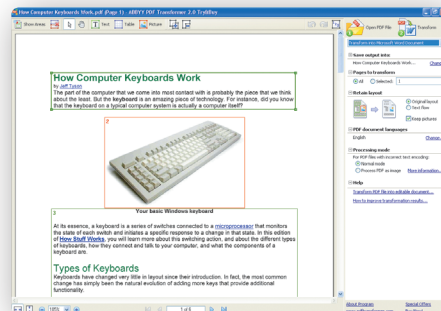
Počecemo od kreiranja PDF dokumenata. ABBYY PDF Transformer kao ulazne formate može da iskoristi Word, Excel i PowerPoint dokumente, ali i tekstualne i HTML fajlove. Ovaj deo programa napravljen je u vidu čarobnjaka, tako da se korisniku ne daje mnogo prostora za podešavanje parametara. Ipak, ono što je ponuđeno

zadovoljiće najveći broj korisnika programa. Od podešavanja je dozvoljeno samo grubo određivanje kolika veličina izlaznog PDF fajla treba da bude. Mala veličina predviđena je za dokumente koji će se pregledati samo na računaru i koji nisu namenjeni za štampanje. Srednja veličina izlaznog fajla može se iskoristiti i za štampanje dokumenata za kućnu ili poslovnu upotrebu, a najveću veličinu treba odabrati kada se dokument priprema za profesionalnu štampu. Na žalost, ne daju se nikakve dodatne informacije o tehničkim specifikacijama ovako kreiranih PDF dokumenata, ali treba imati na umu kome je ovaj program namenjen pa to nećemo uzimati za zlo.

Sa druge strane postoje dve opcije koje su veoma bitne za poslovne (ili privatne) korisnike, a odnose se na zaštićenost i kompatibilnost dokumenata. Zaštićenost dokumenata se ostvaruje na dva načina, zahtevanjem unosa šifre prilikom startovanja dokumenata i zabranom štampanja, editovanja ili kopiranja teksta i slika iz zaštićenog dokumenta. Pošto su PDF standardi vremenom menjani, korišćenjem opcije koja se odnosi na kompatibilnost može se odrediti sa kojom će se najstarijom verzijom Adobe Readera moći otvoriti generisani PDF dokument.

Specijalnost ovog programa je što on veoma lako može da bilo koji PDF dokument pretvori u neki od lakše editabilnih formata. Izlazni formati u kome se mogu naći dokumenti su DOC, RTF, XLS, HTML, TXT ali i PDF format koji se može pretraživati. Uz to prilikom instalacije programa vrši se integracija i sa MS Word i Excel programima iz kojih se nakon toga mogu direktno otvarati PDF fajlovi. Ovaj deo programa predstavlja neku vrstu automatizovane OCR mašine. Isti posao moguće je odraditi i pomoću FineReader-a, ali ovde je to urađeno u stilu vizarda, tako da je posao drastično olakšan za korisnike.

Iako ne nudi sve mogućnosti kao i "veliki brat" PDF Transformer će samostalno obaviti gomilu posla. Treba naglasiti da se sa njim ne mogu ukloniti zaštite koje su postavljene na PDF dokumente jer se zaštićeni dokument neće ni otvoriti ukoliko se ne unese pravilna šifra za otključavanje. Sam proces prepoznavanja teksta i grafičkih elemenata u PDF dokumentu može se vršiti sa očuvanjem izgleda dokumenta, ili u čisto




tekstualnom obliku. Pošto se radi o klasičnom OCR-ovanju od esencijalne je važnosti da se pravilno odabere jezik na kom je dokument napisan. Među četrdesetak podržanih jezika nalaze se i srpska cirilica i latinica, tako da sa našim dokumentima ne bi trebalo da bude problema.

Pošto se radi o digitalnim dokumentima njihov kvalitet je obično na mnogo višem nivou nego kada se vrši OCR-ovanje papirnih dokumenata. Zato je i kvalitet ovog procesa odličan, mogli bismo reći da je tačnost ovog procesa gotovo stoprocentna. Međutim, proizvođač se potrudio da ponudi i još veći kvalitativni mod tako što nudi da se PDF dokument pretvori u sliku pre nego što se krene sa procesom konverzije (navodno je ovakav način OCR-ovanja precizniji i kvalitetniji). Na test dokumentima koje smo mi probali nije bilo potrebe za takvom vrstom konverzije, jer je i direktno iz PDF-a tačnost prepoznavanja bila na gotovo nepogrešiva.

Čini nam se da je ABBYY sa ovim programom pogodio pravo u metu, bar što se tiče poslovnih korisnika gde je u optičaru veliki broj PDF dokumenata. Jednostavnost rada i generalno veoma dobri rezultati koji se postižu su prava preporuka za njegovu korišćenje. Sada više dilema ne postoji - dokumenti se mogu čuvati samo u PDF formatu jer se zna da se kad god je to potrebno mogu ponovo pretvoriti u neki drugi editabilni format.

Branislav Bubanja



### PDF TRANSFORMER

Verzija	2.0 PRO
Proizvođač	ABBYY
Veličina	54 MB
Cena	89 €

[WWW.MFC.CO.YU](http://www.mfc.co.yu)