KOFETKE.

Iz mape PDF-jev v uporabno preglednico

Kaj se izgubi med skeni, izvlečenim besedilom in strukturiranimi podatki?



DOKUMENTI IN PREGLEDNICE

Delovni list za ekstrakcijo s sklici na strani in obravnavo izjem.

BRALNA IZDAJA

september 2026
Lokalna izdaja · september 2026

Viri preverjeni 23. septembra 2026.

Mapa PDF-jev je videti urejena, ker ima vsaka datoteka stalno postavitev strani. To še ne pomeni, da so podatki v njej pripravljeni za preglednico. En PDF lahko vsebuje besedilo, ki ga je mogoče izbrati. Drugi je lahko sken s skrito plastjo OCR. Tretji ima lahko vidno tabelo, katere stolpce predstavljajo samo narisane črte. Četrti je lahko fotografija obrazca, shranjena v ovojnici PDF.

Napaka je, če ekstrakcijo razumemo kot eno samo pretvorbo. Gre za verigo predstavitev: izvirna stran, upodobljena slika, prepoznano besedilo, zaznana postavitev, strukturirane vrstice in celice ter delovna preglednica. Na vsakem koraku se lahko informacija izgubi ali preuredi. Razumen cilj ni »ustvari popoln CSV«. Cilj je »ustvari tabelo, ki ohranja pot nazaj do strani in naredi vprašljive vrednosti vidne«.

Kongresna knjižnica ZDA opozarja, da PDF-ji iz skeniranih strani ne podpirajo nujno indeksiranja besedila, medtem ko je lahko tudi besedilo iz PDF-ja z znaki slabo izvlečeno, kadar notranja zgradba besedila ni čisto preslikana v Unicode.[1] Datoteka, ki je na zaslonu berljiva, lahko zato pri kopiranju v sistem ustvari pokvarjeno besedilo. Pred ekstrakcijo ugotovite, kakšne vrste PDF imate.

Pet plasti, pet različnih vprašanj

Plast	Vprašanje	Kaj lahko gre narobe
Izvirni dokument	Kaj je poslal pošiljatelj ali arhiv?	Manjkajoče strani, neberljiv sken, spremenjena datoteka
Slika strani	Kaj je vidno na strani 7?	Bledi žigi, zasuk, rokopis, odrezan rob
Besedilna plast ali OCR	Kateri znaki so bili prepoznani?	**O** namesto **0**, razdeljene besede, izpuščen minus
Postavitev	Katero besedilo sodi v katero vrstico in stolpec?	Dva stolpca se bereta kot eden, noga strani postane vrstica
Delovna tabela	Katere vrednosti je mogoče izračunati ali filtrirati?	Napačen tip, podvojena glava, izmišljena prazna vrednost

Tabela je nov delovni objekt, ne prosojna kopija strani. Potrebuje svoj identifikator, datum nastanka in povezavo do izvirnika. Model izvora W3C izpeljavo opiše kot pretvorbo ene entitete v drugo.[6] Za uporabo te zamisli ne potrebujete izvedbe semantičnega spleta. Ime datoteke, zgoščena vrednost dokumenta, številka strani in identifikator vrstice omogočijo kasnejši popravek.

Začnite z vhodnim registrom. Za vsako datoteko zabeležite stabilni identifikator dokumenta, prejeto ime datoteke, zgoščeno vrednost, število strani, mesto izvora, datum prejema, navidezni tip in omejitev dostopa. Zgoščena vrednost preverja identiteto, ne resničnosti: pove, ali so se bajti spremenili, ne pa, ali so račun, poročilo ali sken pravilni.

```
Izvirnik PDF
  |
  +--> ohranjena slika strani
  |
  +--> ohranjeno besedilo/OCR
  |
  +--> pregled postavitve
  |
  +--> delovna tabela + sklici na strani
             |
             +--> čakalna vrsta izjem
```

Če ob tabeli hranite sliko strani in izvlečeno besedilo, rešite običajno, vendar drago vprašanje: »Od kod je prišla ta celica?« Če je odgovor samo »iz PDF-ja«, mora pregledovalec znova iskati po celi mapi. Če je odgovor **DOC-014 p. 7, tabela 2, vrstica 5**, se pregled začne na pravem mestu.

Ekstrakcija besedila ni ekstrakcija tabele

Digitalno rojen PDF ima lahko izbirno besedilo v smiselnem vrstnem redu branja. Lahko pa shranjuje posamezne glife, postavljene na strani. Kopiranje take vsebine lahko preuredi stolpce, združi oznake z vrednostmi ali pomišljaj spremeni v

minus. Skenirani PDF morda sploh nima besedilne plasti. OCR doda strojno berljivo besedilo, vendar je to interpretacija pikslov, ne priča, ki je prebrala stran.

Ameriški nacionalni arhiv navaja, da lahko vdelano besedilo OCR izboljša dostop, hkrati pa sprejema, da je lahko nepopravljeno; v svojem okviru arhivskega prenosa zahteva tudi, da zapis PDF po vsebini in videzu ostane enak izvornemu dokumentu.[2] To je koristno razlikovanje za poslovni delovni tok. Izvirno stran ohranite nedotaknjeno. OCR naj pomaga pri iskanju in predizpolnjevanju polj, vendar naj tiho ne zamenja vidne evidence.

Pogosta napaka je, da se obdela OCR na vsaki strani, nato pa se navadno besedilo obravnava kot avtoritativni vhod. Tako se izgubijo vidni dokazi, kot so prečrtana vrstica, polje za podpis, opomba pod tabelo, žig ali meja stolpca. Druga napaka je zavračanje OCR zato, ker ni popoln. S tem zavržete uporaben pripomoček za iskanje. Praktična pot je vmes: OCR uporabite kot izpeljanko, ohranite izvirnik in negotove vrednosti usmerite v pregled.

Tabele potrebujejo zgradbo, ne le besed

Tabela ima odnose. Število v celici je lahko količina, cena na enoto, davek, seštevek, referenčna številka ali noga strani, odvisno od stolpca in vrstice. Če programska oprema zazna besedilo, ne pa teh odnosov, še ni naredila uporabne tabele.

Storitve za ekstrakcijo to razliko izpostavljajo. Microsoftova dokumentacija za postavitev navaja, da lahko izvlečene celice tabele nosijo indekse vrstic in stolpcev ter koordinate omejitvenega poligona; opozarja tudi, da lahko omejitvena območja slik in tabel pri navedeni različici storitve izključujejo pripadajoče napise in opombe.[3] Amazon Textract dokumentira tabele kot bloke z odnosi med celicami, združenimi celicami, naslovom in nogo.[4] To so opisi produktnega vmesnika, ne zagotovila, da bo določen PDF pravilno interpretiran. Pokažejo, zakaj mora delovni tok za pregled ohraniti besedilo in sklic na postavitev.

Uporabno vprašanje ni »ali je orodje vrnilo tabelo?«. Je »ali je pravo vrednost postavilo v pravo polje s pravim dokazom strani?« Vidno urejen rezultat lahko še vedno premakne znesek za en stolpec levo. Združena glava se lahko ponovi kot podatek. Nadaljevalna stran lahko izgubi naslov, ki pojasnjuje spodnja števila.

Za majhen delovni tok uporabite delovno tabelo s stolpcem za izvor, namesto da bi izvor skrivali v ločenem sistemu:

document_id

DOC-014

stran	7
table_ref	T2
source_row	5
datum	2026-08-31
dobavitelj	Vektor d.o.o.
znesek	146,40
valuta	EUR
zaupanje / preverjanje	vizualno preverjeno
izjema	ne

DOC-014

stran	7
table_ref	T2
source_row	6
datum	2026-08-31
dobavitelj	Vektor d.o.o.
znesek	18,00
valuta	EUR
zaupanje / preverjanje	nizko zaupanje OCR
izjema	potrdi **8** ali **B**

table_ref in **source_row** sta lokalna sklica, ki ju ustvarite sami. Ne trdita, da izvirni PDF vsebuje ta identifikatorja. Pregledovalec ju lahko s številko strani uporabi za iskanje ustreznega vidnega

območja. Če orodje za ekstrakcijo zagotovi lastne koordinate ali identifikatorje celic, jih ohranite kot tehnični dokaz, vendar ne zahtevajte, da jih običajen pregledovalec razvozla, preden vidi stran.

Izmišljeni primer: 48 izpiskov dobaviteljev

Oglejmo si izmišljeno podjetje Kovač Workshop. Knjigovodja prejme 48 mesečnih izpiskov dobaviteljev v PDF-jih. Lastnik želi pred pregledom računovodje tabelo z datumom izpiska, imenom dobavitelja, sklicem, neto zneskom, davkom, bruto zneskom in valuto. To je priprava, ne knjiženje ali oddajanje česar koli.

Vhodni register dodeli **STAT-2026-001** do **STAT-2026-048**. Za vsak izvirnik zabeleži prejeto ime datoteke in zgoščeno vrednost. Prvi prehod prepozna 31 digitalno rojenih PDF-jev, 14 skenov s plastjo OCR in tri skene brez uporabnega besedila. Delovna tabela ne pretvarja, da so vse vrstice nastale iz iste vrste dokazov. V polju **extraction_method** piše **native_text**, **OCR** ali **manual_transcription**, v polju **review_state** pa **unreviewed**, **checked** ali **exception**.

Pri izpisku z običajno mrežo prvi prehod predlaga vrstice. Pregledovalec odpre izvorno stran ob predlagani vrstici. Preveri ime dobavitelja, datum izpiska, valuto in seštevek. Preveri tudi, ali tabela vsebuje preneseno stanje ali skupni seštevek; oboje je videti kot znesek transakcije, vendar pomeni nekaj drugega. Podpis ali žig sta lahko pomembna za poslovni postopek, nista pa samodejno polje za računovodsko tabelo.

Tri strani brez uporabnega besedila se ne silijo skozi ugibani rezultat OCR. Vstopijo v čakalno vrsto izjem s slikami strani in razlogom **sken je preveč bled za zanesljiv OCR**. Ta zapis ekipi pove, zakaj je v tabeli vrzel in kaj potrebuje človeško pozornost.

Primer napake: aritmetika je pravilna za napačne vrstice

Predpostavimo, da ima en izpisek dvovrstično ime dobavitelja in stolpec zneskov, poravnan na desno. Ekstraktor nadaljevalno vrstico prebere kot novo vrstico. Nato prvi znesek pripne napačnemu dobavitelju. Bruto seštevek na dnu se še vedno ujema z vsoto izvlečenih zneskov, zato preprost aritmetični pregled uspe.

To je nevarna lažna gotovost. Uskladitev lahko najde manjkajoče ali podvojene zneske, ne more pa dokazati pomenskega para dobavitelja in zneska. Potreben je vzorčni vizualni pregled različnih postavitev. V vzorec vključite nerodne strani: zasukane skene, združene celice, navidez prazne vrstice, večstranske tabele, žige, ročne popravke in negativne zneske.

Druga napaka nastane, kadar motor OCR v referenčni številki prebere **8** kot **B**. Če tabela to polje pretvori v število, lahko napaka postane prazna ali spremenjena vrednost brez sledi. Identifikatorje hranite kot besedilo, ohranite vodilne ničle in označite nepričakovane znake. Dobesedno ujemanje je pomembno za številke računov, bančne reference, datume in kode. Iskanje po podobnosti lahko pomaga najti verjetne zapise; identifikatorja ne sme tiho »popraviti«.

Obravnava izjem je del zasnove

Čista tabela brez stolpca za izjeme pogosto pomeni, da je bila negotovost zavržena, ne rešena. Dodajte register izjem z identifikatorjem izvirnega dokumenta in sklicem na stran. Uporabne kategorije so:

- ni besedilne plasti ali OCR ni uspel;
- stran manjka, je zamegljena, zasukana ali delno odrezana;
- tabela se nadaljuje na več straneh;
- glava ali noga je morda postala podatkovna vrstica;
- združene celice ali večvrstična oznaka zahtevajo razlago;
- znesek, predznak, datum, valuta ali identifikator je dvoumen;
- navidez podvojena vrstica potrebuje potrditev;
- dokument je zunaj dogovorjenega obsega ali meje dostopa.

Vsaka izjema potrebuje lastnika, stanje in opombo o razrešitvi. »Razrešeno« mora povedati, kaj se je spremenilo in kdo je preveril, na primer »potrjeno glede na **STAT-2026-014 p. 3**; pregledovalec je OCR **B** popravil v **8**«. Izvirnega rezultata OCR ne zamenjajte, ne da bi ga ohranili. Sicer poznejši pregledovalec vidi le popravljeno tabelo in ne more ugotoviti, zakaj je bil popravek narejen.

Tehnične smernice Kongresne knjižnice za en program digitalizacije zahtevajo eno datoteko OCR na sliko strani in podatke o zaupanju navajajo kot uporaben dodatni element.**[5]** Majhnemu projektu PDF-v-tabelo te specifikacije ni treba kopirati.

Lekcija zasnove je ožja: povezava na ravni strani in podatki o negotovosti so uporabni, ker je rezultat ekstrakcije lažje preveriti, kadar ostane povezan s stranjo.

Strnjena metoda vrednotenja

Pred uporabo tabele za odločitev izvedite tri preglede.

1. **Pregled popolnosti.** Primerjajte število strani v registru s številom obdelanih strani in navedite vsako izključeno stran. Samo število vrstic ne dokazuje, da so bile upoštevane vse strani.
2. **Pregled polj.** Vzemite stratificiran vzorec različnih vrst dokumentov in postavitev strani. Pri vsaki vzorčni vrstici primerjajte datum, identifikator, znesek, valuto in dobavitelja z izvorno stranjo. Zabeležite napake po poljih, ne samo rezultat uspešno/neuspešno.
3. **Pregled uskladitve.** Kjer dokument navaja seštevek, s pomočjo orodja izračunajte vsoto ustreznih izvlečenih vrstic in jo primerjajte s tem seštevkom. Neujemanje obravnavajte kot izjemo. Ujemanje obravnavajte kot uporaben dokaz, ne kot dokaz, da je vsaka vrstica pravilno pripisana.

Za vsote uporabite orodja in ohranite vhodne podatke izračuna. Če na primer pet preverjenih vrstic vsebuje 146,40, 18,00, -32,50, 90,10 in 7,00 EUR, je izračunani seštevek 229,00 EUR. To število mora izpeljati preglednica ali skripta, ne računanje na pamet. Še vedno ne more potrditi, da 18,00 pripada pravemu dobavitelju. Vizualni pregled vrstice pokriva drugo vrsto napake.

Delovna tabela si zaupanje prisluži vrstico za vrstico: vsaka vrednost ima izvorno stran, vsak izpeljani seštevek je mogoče ponoviti, vsaka vprašljiva ekstrakcija pa ostane vidna, dokler je človek ne razreši. To je počasneje od izvoza z enim klikom. Je pa razlika med uporabno pripravljalno datoteko in lepo urejeno zbirko napak.

Viri

Številke virov veljajo za ta vodnik. Povezave vodijo do izvirnih dokumentov. Pred uporabo preverite besedilo, pogoje uporabe in razpoložljivost storitev.

[1] **PDF (Portable Document Format) Family**
https://www.loc.gov/preservation/digital/formats/fdd/fdd000030.shtml

[2] **Transfer Instructions for Permanent Electronic Records in PDF format**
https://www.archives.gov/records-mgmt/initiatives/pdf-records.html

[3] **Document layout analysis - Document Intelligence**
https://learn.microsoft.com/en-us/azure/ai-services/document-intelligence/prebuilt/layout?view=doc-intel-4.0.0

[4] **Tables - Amazon Textract**
https://docs.aws.amazon.com/textract/latest/dg/how-it-works-tables.html

[5] **National Digital Newspaper Program 2024 Technical Guidelines for Applicants**
https://www.loc.gov/ndnp/guidelines/NDNP_202426TechNotes.pdf

[6] **PROV-O: The PROV Ontology**
https://www.w3.org/TR/prov-o/