

Konferencija za korisnike Transkribusa

Karmen Levanić



Polaznici Transkribus User Conference 2017 ©Ingrid Reinhardt CC BY-NC-ND 4.0

U Beču je 2. i 3. studenog 2017. održana konferencija za korisnike Transkribusa. Transkribus je sveobuhvatna platforma za automatsko prepoznavanje, transkripciju i pretraživanje povijesnih dokumenata, koja se razvija u sklopu READ projekta (*Recognition and Enrichment of Archival Documents* – Prepoznavanje i obogaćivanje arhivskih dokumenata). Glavni je cilj Transkribusa pružanje potpore korisnicima, koji se bave transkripcijom rukom pisanih ili tiskanih dokumenata – istraživačima, arhivistima, knjižničarima, ali i široj javnosti zainteresiranoj za proučavanje povijesnih dokumenata. Transkribus nudi brojne alate za automatsku obradu dokumenata, npr. prepoznavanje rukom pisanog teksta (HTR, *Handwritten Text Recognition*), OCR, pretraživanje ključnih riječi i indeksiranje (*Keyword Searching and Indexing*), povezivanje slike i teksta (*Text2Image Matching*), prepoznavanje pisca (*Writer Identification*). Korištenje Transkribusa i svih njegovih alata i aplikacija je besplatno, budući da ga financira Europska komisija kao dio READ projekta.

Konferencija se održavala u zgradi Tehničkog sveučilišta u Beču, iz koje se pružao prekrasan pogled na vizure Beča, okupanog kasnojesenjim suncem. Na početku konferencije prisutnima se riječima dobrodošlice obratio dr. sc. Günther Mühlberger sa Sveučilišta u Innsbrucku, koordinator READ projekta. Predstavio je Transkribus platformu, a zatim je govorio o tome gdje se trenutno nalaze i što žele postići

u budućnosti. Cilj je projekta da se, prikupljanjem što većeg broja dokumenata i transkripcijom, Transkribus isprofilira za prepoznavanje raznih pisama i rukopisa te da postane vodeća platforma za transkripciju i pretraživanje rukom pisanih i tiskanih tekstova.

U nastavku su o Transkribusu govorili njegovi korisnici te se tijekom dva dana konferencije predstavilo 11 izlagača iz različitih arhiva, sveučilišta i istraživačkih grupa iz Austrije, Njemačke, Nizozemske, Španjolske i Srbije. Oni su u svojim prezentacijama prikazali kako koriste Transkribus i koje su rezultate postigli njegovom primjenom u obradi raznih vrsta dokumenata, od srednjovjekovnih isprava i inkunabula, preko novina iz 18. st. te pisama i matičnih knjiga iz 19. st., do rukopisa Samuela Becketta. Time su dali uvid u mogućnosti Transkribusa, čijom primjenom nastaje ne samo automatski prepisan, nego i u potpunosti pretraživ tekst.

Program je potom nastavljen u dvije paralelne sesije – za početnike i za napredne korisnike Transkribusa. Novim su korisnicima o prepoznavanju i pretraživanju teksta te indeksiranju govorili matematičar, Roger Labahn sa Sveučilišta u Rostocku i Alejandro Tosselli s Politehničkog sveučilišta u Valenciji. Naprednim je korisnicima predstavljen novi, poboljšani alat *Text2Image*, koji se koristi postojećim digitaliziranim dokumentima i transkripcijama za daljnje automatsko prepoznavanje teksta, a govorili su i o

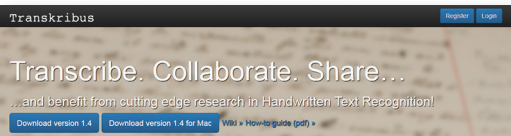
potrebi povezivanja humanističkih i kompjuterskih znanosti te o napretku, koji se postiže natjecanjima.

Usljedila je zajednička sesija, gdje su prikazani dokumenti, koje su poslali polaznici Konferencije, a Günther Mühlberger i ostali stručnjaci iz Transkribus tima komentirali su primjenu Transkribusa u obradi tih dokumenata i odgovarali na pitanja korisnika.

U poslijepodnevnom dijelu ponovno su održane dvije paralelne radionice, za nove i za napredne korisnike. Novi su se korisnici upoznali s osnovama korištenja Transkribusa, kako instalirati program i kreirati zbirku isprava, a zatim pristupiti unošenju podataka, potrebnih da bi se program mogao isprofilirati za daljnje automatsko prepoznavanje teksta. Za to je potrebno tekst prvo razdijeliti u dijelove i redove (segmentacija teksta) i onda unositi transkripciju po segmentima. Na drugoj radionici iskusniji korisnici upoznavali su se s naprednijim tehnikama rada i alatima, koje Transkribus nudi.

Program je završen predstavljanjem demo verzije Transkribusovog web sučelja, koje je još u izradi, a zamišljeno je kao jednostavnija inačica programa, koja se brže i lakše usvaja te je predviđena za najširi krug korisnika. Na kraju prvoga dana priređena je večera na kojoj su se sudionici konferencije mogli ugodno družiti i izmijeniti svoja iskustva u neformalnoj atmosferi, a tijekom večere organizirana je tombola na kojoj je podijeljen uređaj *ScanTent*, koji služi za digitalizaciju dokumenata putem mobitela.

Drugog je dana konferencija nastavljena izvještajima korisnika, a potom se održala „tržnica“ alata, koji se razvijaju u sklopu READ projekta. Ta je „tržnica“ zamišljena tako da je svaki stručnjak za pojedino područje iz Transkribusovog tima imao svoj „štanđ“, a svaki je sudionik mogao ići od stola do stola i postavljati pitanja. Predstavljeni su alati za analizu izgleda stranice i automatsku segmentaciju (*Layout Analysis*), za povezivanje slike i teksta



Scholars
Are you transcribing historical documents? Handwritten or printed. From the middle ages or from the 20th century? Would you like to do this in a highly standardized, flexible and repeatable way? And do you appreciate to get support from automated tools such as Handwritten Text Recognition and Layout Analysis?

Archives
Are you responsible for large collections of handwritten and printed documents? Do you believe that digitization opens the way to realise new opportunities to access, search and explore archival material? And are you open to involve transcribing scholars and volunteers so that they can work with these documents in an effective way - producing data which can also be integrated in your repository?

Volunteers
Are historical letters, postcards, manuscripts or medieval documents fascinating for you? Do you enjoy deciphering handwriting - this wonderful feeling when you can read something which may be hidden to most other people? And do you believe that everyone can make a valuable contribution to scholarship and science?

Scientists
Are you a computer scientist and working in the fields of computer vision, document analysis, pattern recognition, natural language processing or a related field? You are seeking interesting documents from 1000 years of handwriting, printing and pastpapers? And you would really enjoy to get some reference data in a well-acknowledged format, such as a PAGE?

Humanities Scholars
TRANSKRIBUS enables you to:
- register in the platform
- develop an expert tool specifically designed for your needs
- upload your own documents
- manage your own private collection (no one else has access to your documents)
- segment the images into books, manuscripts and records with the support of layout analysis tools
- use the tool with images which increased the value of your transcription significantly (usually you cannot transcribe text without images)

TRANSKRIBUS enables you to:
- transcribe text in any language and with any character set (not your own virtual keyboard)
- export your documents at every time in several formats such as TEI, PDF, PPT, XML
- Now, this sounds interesting, but it starts to get really exciting if you consider that:
- you can have your properly transcribed e.g. 100 images you upload to the platform and train an HTR engine from the Computational Intelligence Technology Lab (CITLAB) of the University of Roehampton on your documents and
- you will be able to transcribe further images of your documents with the support of automatically produced transcription text

Get started!
1. Register
2. Download Transkribus
3. Try out some test documents with automatically produced full-text in the TranskribusCloud collection

Archives and Libraries
TRANSKRIBUS enables you to:
- safely the needs of your users who are working with your documents by keeping full control over your documents
- TRANSKRIBUS humanities scholars will get one of the most advanced transcription tools currently available. But still, current tools: Transcriptions will not get "lost" on the private computers of your users, but will also be available for enriching your document collection and for getting integrated into your repositories in a standardized and seamless way
- organise cross-sourcing projects with a minimum of effort. Based on the experiences of one of the most successful cross-sourcing projects "Transcribe Bentham" your cross-users will work with a dedicated cross-sourcing interface which was developed by the University College London in close cooperation with the University London Computer Centre and which is also part of the TRANSKRIBUS platform.

Now, this sounds interesting, but it starts to get really exciting if you consider that:
- every user collaboration takes properly transcribed at least 100 pages will be possible to train a cutting edge HTR engine developed by the Computational Intelligence Technology Lab (CITLAB) of the University of Roehampton and to recognise the rest of the documents in an automated way. The results of the automated transcription are amazing, and the good news is that the more data are available for training the HTR engine, the more likely it is possible to improve the recognition accuracy.

Interested?
1. Request more information - we are happy to answer your questions
2. Register in the platform
3. Download TRANSKRIBUS
4. Try out the tool - so you will get an impression about the working environment of your collaborators and volunteers
5. Upload some test documents (or just some images) to your private collection to make first experiments with your own documents (don't worry, no one else will be able to access them)

Volunteers
TRANSKRIBUS enables you:
- to work with your own documents! In this case you will be able to:
- use the same tool as "professional scholars" - your contributions will therefore be of the same high quality which is required for serious science
- upload your own documents (or those from your grandparents, relatives, etc.) to the platform
- keep full control on the documents (don't worry if you who decides who has access to the documents)

TRANSKRIBUS enables you to:
- involve other users who really support you in the transcription of documents or even to engage a professional transcription service who can do the transcription for you
- export your documents not only as HTR (crowd text) but also as PAGE (with the image in the foreground and the text in the background)

Get started!
1. Register
2. Download TRANSKRIBUS and try it out

Scientists
TRANSKRIBUS will enable you to:
- register in the platform and stay informed about interesting transcription projects
- upload your own images and train them against core tools developed in the PPT project transcription
- use the TRANSKRIBUS tool for accessing your own data with regions of interest and to store the results as PAGE files (compatible to ALETHRA from USA)
- to communicate with interested humanities scholars and archives on the basis of clearly defined reference data and benchmarks (work on the same documents, store separate versions of the same document)

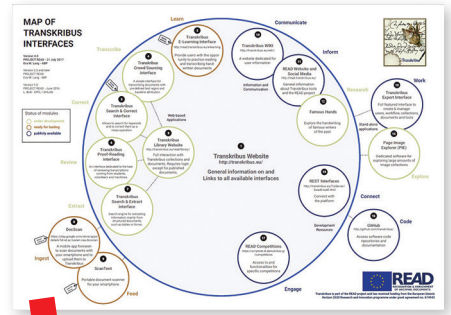
But what is really exciting is that TRANSKRIBUS is also a way to have a beautiful interface for your research results. It enables you:
- to explore your tool in a well-known environment (Citrus/WindowsServer) with clearly defined input-output formats (PAGE in PAGE out)
- to gather feedback from a variety of users such as humanities scholars and volunteers
- to keep your tools running without having to care of the server (server, document management, user management, logging, bandwidth, support, communication...)

Get started!
1. Contact us
2. Register in the platform
3. Download TRANSKRIBUS and try it out

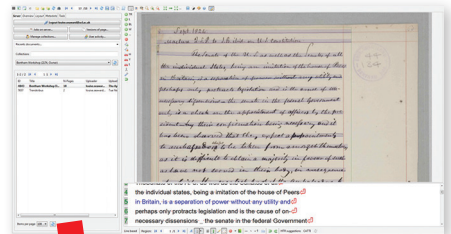
projekta i završnim riječima konferenciju priveo kraju.

Iskustva korisnika pokazala su vrlo široku primjenu Transkribusa za automatsku transkripciju i pretraživanje raznih vrsta tekstova. U teoriji, on bi se mogao koristiti za čitanje svih vrsta pisama i rukopisa, no da bi program radio automatski potrebno ga je prvo naučiti da sam segmentira tekst i da ga zatim transkribira. To znači da prvo netko treba ručno podijeliti tekst u segmente, a zatim ručno unijeti transkripciju za oko 20.000 riječi, odnosno 100 stranica teksta. Kad se unese osnovni skup podataka, obavještava se Transkribus tim, pokreće se postupak učenja i kreira modul, koji omogućuje računalo da dalje automatski segmentira i transkribira tekst toga tipa. Potrebni su opsežni pripremni radovi, a nakon što računalo završi posao, potrebno je sve ručno prekontrolirati, međutim, u pojedinim primjerima Transkribus je došao do stupnja pogreške od 5 %, što je sjajan rezultat.

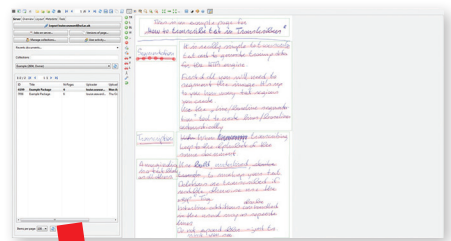
Uvjerena sam da bi i kod nas Transkribus pronašao svoju primjenu. Kako pripremni radovi za treniranje programa predstavljaju prevelik posao za jednu osobu, bilo bi dobro pronaći zajednički interes i stvoriti tim ljudi, koji će raditi na tome. Jednu od mogućnosti primjene Transkribusa vidim u transkripciji isprava. Budući da u kontinentalnoj Hrvatskoj baštinske ustanove imaju slične kasno srednjovjekovne i rano novovjekovne isprave, koje dolaze iz istih izvora (kraljevska i banska kancelarija, kaptoli), učenje Transkribusa na automatsku transkripciju takve vrste dokumenata bilo bi od koristi za sve koji ih posjeduju. Tako bismo zajedničkim snagama, uz pomoć kompjuterskih i razvojnih stručnjaka iz Transkribus tima, mogli razviti naš zajednički modul za transkripciju isprava i na taj način to vrijedno gradivo učiniti dostupnim i pretraživim širokom krugu istraživača. ■



Mapa Transkribus sučelja © READ



Transkribus sučelje za rad korisnika sustava



Transkribus „paket primjera“

INFO

Platforma Transkribus dostupna je na: <https://transkribus.eu/Transkribus/>

Transkribus YouTube program dostupan je na: <https://www.youtube.com/channel/UC-txVgM31rDTGIBnH-zpPjA/featured>

Program i materijali s Transkribus User Conference 2017 dostupni su na: <https://read.transkribus.eu/transkribus-user-conference/>

Blog READ Projekta dostupan je na: <https://read.transkribus.eu/category/success-stories/>