



РЕФЕРЕНЦЕ УСТАНОВЕ/ОРГАНИЗАЦИЈЕ:

Универзитетска библиотека „Светозар Марковић“ (УБСМ) је централна академска библиотека Универзитета у Београду и матична високошколска библиотека на територији централне Србије. УБСМ у свом фонду поседује око 2 милиона јединица. Данас, УБСМ је савремени центар информационог друштва, друштвено препознати лидер у области дигитализације и институција са великим искуством у међународним пројектима. Једна од централних активности УБСМ јесте едукација библиотечке, студентске и истраживачке заједнице. УБСМ има широку скалу едукативних активности – од едукације истраживача о програмима, методама и ресурсима, преко едукације студената докторских студија, уз подршку Министарства науке, технолошког развоја и иновација, до едукација студената, библиотекара и истраживача из области дигиталне хуманистике, отворене науке, језичких технологија, примене вештачке интелигенције и других актуелних тема. УБСМ у редовима запослених има неколико међународно признатих експерата из ових области који су чланови међународних стручних тела, радних група, програмских одбора.

Друштво за језичке ресурсе и технологије (JePTex) је друштво основано ради остваривања циљева у области промоције, популаризације и унапређивања свих грана језичких технологија на научном, стручном и практичном нивоу. Чланови друштва су међународно препознати стручњаци, са огромним искуством у међународним пројектима, врхунским научним референцама и великим искуством у едукацији студентске и истраживачке заједнице. JePTex је пионир у креирању језичких технологија за српски језик и данас располаже највећим ресурсима, апликацијама и моделима за српски језик.

И УБСМ и JePTex су претходне две године успешно имплементирали петодневне радионице за студенте основних, мастер и докторских студија из области дигиталне хуманистике у Научно-образовном културном центру „Вук Караџић“ у Тршићу.

РЕАЛИЗАТОРИ ПРОГРАМА:

Др Василије Милновић – Руководилац Центра за науку, Универзитетска библиотека „Светозар Марковић“, Булевар краља Александра 71, 11120 Београд. Руководилац неколико пројеката из области дигиталне хуманистике, српски представник у *European Open Science Cloud Steering Board*, Брисел и члан Радне групе LIBER (Association of European Research Libraries) *Digital Scholarship and Digital Cultural Heritage Collections Working Group*.

Проф. др Ранка Станковић – шеф Рачунарског центра Рударско-геолошког факултета (Ђушина 7, 11120 Београд) и професор на Катедри за примењену математику и информатику. Потпредседник је Друштва за језичке ресурсе и технологије. Ментор стварања првог генеративног модела за српски језик *OPAO GPT2*. Главни и одговорни уредник часописа за дигиталну хуманистику „Инфотека“.

Проф. др Цветана Крстев – професор Филолошког факултета Универзитета у Београду у пензији и суоснивач Друштва за језичке ресурсе и технологије, координатор српског учешћа у COST Акцији *Distant Reading*, дугогодишњи уредник часописа за дигиталну хуманистику „Инфотека“ и аутор неколико електронских речника и корпуса српског језика.

Проф. др Душко Витас – оснивач и дугогодишњи шеф Катедре за рачунарство и информатику, у пензији, Математички факултет Универзитета у Београду, оснивач Друштва за језичке ресурсе и технологије и саветник COST Акције *Distant Reading*, аутор СрпКор корпуса и неколико електронских речника и корпуса српског језика. Носилац ордена Републике Француске.

Др Александра Трговац – библиотекар саветник Универзитетске библиотеке „Светозар Марковић“ (Булевар краља Александра 71, 11120 Београд), шеф каталога, члан Програмског одбора LIBER (Association of European Research Libraries) и члан Радне групе *Digital Scholarship and Digital Cultural Heritage Collections Working Group* исте асоцијације. Координатор неколико пројеката из области дигиталне хуманистике и технички уредник часописа за дигиталну хуманистику „Инфотека“.

ОЧЕКИВАНИ ИСХОДИ: оспособљавање полазника у припреми текстова за корпус/дигиталну колекцију. Предвиђено је да радионице буду радног карактера. Полазници ће се кроз



практичан рад упознати са целом линијом дигитализације: корекција сканираних и прочитаних текстова, аотирање њихове структуре и изгледа, припремање мета-података. Напредним радионицама предвиђа се обучавање полазника за проверу и корекцију аутоматски обележених језичких објеката: имена људи, њихових улога, локација и друго. За ове потребе користиће се алати доступни на: <http://nerbeyond.jerteh.rs/>. Такође, полазници ће се упознати са применом језичких модела, повезивањем са отвореним подацима, електронским речницима, могућностима аутоматизације. Посебан сегмент ће се бавити проналажењем валидних ресурса и информација, као и дигиталних алата и њихова примена на текст. Кроз практичан рад, полазници ће усвојити веома много информација из области креирања језичких технологија, дигиталне хуманистике, али и удаљеног читања и отворене науке. Биће оспособљени да повећају видљивост свог рада, па самим тим и теме својих радова, а то је српска култура и језичко наслеђе у широком смислу.

ОБЛИЦИ РАДА: комбиновање теоријског и практичног рада. Осим теоријских излагања тема наведених у претходној секцији, полазници би имали прилику да припреме један текст, од корекција и обележавања TEI етикета до аутоматске аотације именованих ентитета. Припремљене текстове би полазници користили за повезивање са отвореним подацима, а потом би уследили практични задаци удаљеног читања и проширивања упита за кориснике. Ресурси за практичан рад: srpELTeC корпус, колекција историјских новина - *Претражива* и језички модели развијени у оквиру пројекта ТЕСЛА.

ТРАЈАЊЕ АКТИВНОСТИ И ВРЕМЕНСКИ РАСПОРЕД АКТИВНОСТИ: предавања се организују 5 радних дана, пре подне, у периоду од 10-13 часова са краћим паузама, а радионице пет радних дана, после подне од 16-18 часова.

ДИЗАЈН АКТИВНОСТИ: Структура дизајна активности садржи: ток, ниво, главне поруке (кључне тачке), методе, аудиовизуелна средства, писани материјали, остали материјал и прибор.

Распоред за 5 дана:

Дан 1

Теоријски део (пре подне):

- Важност дигиталне хуманистике данас, резултати досадашњих пројеката, генералне информације о COST Акцији и циљевима дигиталне хуманистике на овом примеру, дигитална хуманистика и отворена наука, питање канона и дигиталне хуманистике.

- процес дигитализације (скенирање, рашчитавање), алатке које се користе. На примеру <https://udaljenocitanje.unilib.rs/> демонстрација разлика слике и рачитаног текста.

Карактеристични проблеми за рашчитавање ћириличног и латиничног писма, проблеми старих и ретких књига, пробелми рашчитавања новина. Аутоматска корекција текста применом језичких модела.



- Значај стандардизације (TEI, пример Вукових пословица, ELTeC-a).

Практичан рад (после подне):

Сваки полазник креира сопствени мини пројекат у ком креће од рашчитавања скенираног документа (<https://github.com/procesaur/digitek>), након чега следе корекције текста и структурне анотације. Примери ће обухватити оба писма.

Дан 2

Теоријски део (пре подне):

- Обрада дигиталних текстова. Сегментација текста на реченице и токенизација. Врсте и значај анотација.
- Аутоматска анотација токена врстама речи и лемама, аутоматска анотација делова текста именованим ентитетима (места, особе, организације, професије,...).
- Текстометрија.
- Напомене о напреднијим врстама анотације: синтакса, директан говор, семантика (значења), стилске фигуре. Неопходност контроле и евалуације.

Практичан рад (после подне):

Контрола ручне анотација структуре документа сагласно TEI препорукама (добро формиран и валидни XML документи); аутоматска анотација врстама речи и лемама (<http://obrada.jerteh.rs/>), именованим ентитетима (<https://ners.jerteh.rs/>). Контрола и евалуација аутоматске анотације (<http://inception.jerteh.rs/>).

Упознавање са корпусима доступним на <https://noske.jerteh.rs/>. Основе CQL језика и демонстрација упита над корпусом srpELTeC. Анализа резултата коришћењем конкорданци, фреквенција, извоз резултата. Инсталација алата TXM.

Дан 3

Теоријски део (пре подне):

- Отворени подаци, упознавање са википодацима о српским романима: основни подаци о роману и његовим издањима, подаци о аутору, о местима радње, главним ликовима и њиховим везама. Ручни и аутоматски унос википодатака.
- Демонстрација рада са Википодацима из пајтон свески.
- Претраживање података језиком SPARQL и визуелизација резултата табеларно, у виду мапе, графа, стабла. Интеграција резултата упита у презентације.

Практичан рад (после подне):



Сваки полазник ће унети по један роман у Википодатке, креирати упите са различитим излазима и интегрисати их са документом (HTML). Повезивање ентитета у тексту са Википодацима.

Дан 4

Теоријски део (пре подне):

- Улога библиотека у дигиталној хуманистици: пружање приступа дигиталним збиркама и базама података, обука и подршка у коришћењу дигиталних технологија за анализу и обраду хуманистичких података; очување дигиталне културне баштине (дигиталних копија старих рукописа и других важних историјских докумената). Улога библиотекара у дигиталној хуманистици на примеру конкретног прикупљања материјала за припрему корпуса. Ресурс: историјске новине из колекције – Претражива.

Практичан рад (после подне):

рад у програму *Transkribus*, напредно претраживање у COBISS-у.

Дан 5

Теоријски део (пре подне):

- Читање у дигиталном окружењу - е-књиге. Текстометријска анализа текста коришћењем алата ТХМ: креирање корпуса, генерисање речника текста, конкорданце и фреквенције CQL образаца, креирање и анализа партиција и подкорпуса, анализа специфичности.

- Шири контекст српског језика у доба револуције вештачке интелигенције (ВИ), могућности и проблеми у вези са обрадом српског језика коришћењем ВИ, посебно у контексту великих језичких модела (попут OpenAI GPT), који тренутно привлаче глобалну пажњу. Статички и динамички језички модели. Иницијатива за развој јавно доступних квалитетних ресурса и алата за обраду српског коришћењем модела ВИ, како би се сачувала позиција српског језика у доба АИ револуције.

Практичан рад (после подне):

- Креирање сопственог малог корпуса текстова и његова текстометријска анализа.

- Коришћење модела за српски: ГПТ, фамилије BERT и других са платформе <https://huggingface.co/jerteh>.

- Инструкције (промптни инжењеринг) на примеру ChatGPT и других модела.

Семинар подразумева личне лап топ уређаје полазника, РРТ презентације, пројектор, интернет, дигиталне колекције. Распоред тема по данима може бити измењен.