Международна конференция „The Digitisation Days“ или новите постижения в технологиите за цифровизация на текст
The Digitization Days International Conference or the last achievements in digitization

Ваня Русева
Vanya Ruseva

РЕЗЮМЕ:
Авторът, началник отдел „Технически ресурси и услуги“ в библиотеката на НБУ, споделя своите впечатления от международната конференция “Digital Access to textual Cultural Heritage” (DATeCH) в рамките на инициативата „Digitisation Days“, организирана от IMPACT Center of Competence in Digitisation и Succeed. Домакин е Националната библиотека на Испания, гр. Мадрид. Събитието среща на едно място библиотекари, компютърни специалисти, изследователи, фирми – разработчици на най-новите, модерни технологични решения в областта на дигитализация на текстове от античността до 19 в.

ABSTRACT:
The author, Head of the Technical Resources and Services Department at the NBU Library, shares her impressions of the International Conference "Digital Access to Textual Cultural Heritage" (DATeCH) in the frame of the "Digitization Days" Initiative, held by IMPACT Center of Competence in Digitization and Succeed. The event has been hosted by the National Library of Spain, in Madrid. The venue opens a forum for discussion between librarians, programmers, researchers and developing companies, offering the newest technological solutions for digitizing text from the Antiquity to the 19th Century.

Антикварна книжарница (за литература и история) - една от многото по централните улички в Мадрид, популярни сред библиофилите.

Реализирани са стотици проекти за опазване на писменото културно наследство и осигуряване на достъп до него. В резултат са натрупани огромни масиви от дигитализирани документи. Благодарение на тези проекти, на целевото финансиране за

тях, технологиите се развиват и продължават своя прогрес в посока бързо и щадящо сканиране на оригиналите, изчистване и коригиране на изображенията, създават се все по-добри платформи за тяхното описание, индексиране и организиране в колекции. Всички тези усилия заслужават адмирации, но фокусът при опазването и достъпността до писмените паметници на човечеството се измества към качеството на текстовете в цифров вид, т.е. до каква степен масивите удовлетворяват информационните и изследователски нужди на учени и практици, при положение че тези дигитални колекции предоставят достъп предимно до изображения, или в най-добрия случай частично, неточно разпознати и коригирани текстове. Това именно е новото предизвикателство пред учени и практици с компетенции в различни области на знанието - да подобряват наличните ресурси, да разработват нови такива за създаване, преобразуване и разпространение на дигитализираните текстове с историческо значение. Това предизвикателство предполага интердисциплинарно сътрудничество, то насърчава организирането на различни инициативи.

Една от тях е проведената на 19 и 20 май 2014 г. в Мадрид международна конференция „The Digitisation Days“. Тя се проведе с финансовата подкрепа на Европейската комисия в Националната библиотека на Испания. Организатори са център IMPACT и европейската инициатива Succeed. IMPACT е център за компетенции в областта на цифровизацията и е създаден в резултат на реализиран едноименен проект. Целта на тази европейска инициатива е да съдейства и разпространява най-добрите технологии и практики. В него участват над 35 организации, сред които и едни от най-големите и значими по отношение на културното наследство европейски национални библиотеки. Проектът Succeed също е европейска инициатива, която подпомага разработването на инструменти и ресурси, насърчава обмена на знания, като съдейства за организиране на изследователски консорциуми и на публично-частни партньорства, за да се реализира в качествено отношение масовата дигитализация на писмените паметници. Организаторите замислят и провеждат конференцията с идеята да бъде място за среща на библиотекари, изследователи и компании с цел да представят нови постижения, да обменят опит и споделят проблеми, да обсъдят бъдещи предизвикателства.
„The Digitisation Days“ се състоя от:

- конференция под наименование DATECH (Digital Access to Textual Cultural Heritage), по време на която в няколко тематично обособени сесии бяха представени резултати от разработки и проекти;
- фирмени презентации на големи, водещи компании като Contenta Technologies, Libnova, i2s Digibook, Technilogica, Digibis и демонстрации на най-модерните им решения;
- представяне на плакати, чрез които колективи от специалисти визуално демонстрираха резултати от техни разработки и експерименти;
- церемония по връчване на награди от инициативата Succeed за най-успешните реализации;
- панелни дискусии относно политиките за дигитализация, права на интелектуална собственост и други горещи теми в сферата на цифровизацията на исторически текстове;
- панелна сесия под надслов „Дигитализацията на културното наследство: Модерна утопия?“ с участието на UNESCO, Europeana, Standford University Library, Europea Commision и Utrecht University.

Парадното стълбище и вход на националната библиотека на Испания

Зала в Кристалния дворец (Crystal Palace), който се намира в Buen Retiro Park в Мадрид.

В тематичната сесията „Анализ на документи, оптично разпознаване и коригиране“ (Document Analysis & OCR) бяха представени разработки на методи за автоматизирано сегментиране на статии и графични обекти от дигитализирани старопечатни вестници, организирани в електронно достъпни колекции; генериране на мета данни в METS/ALTO и други стандартизирани формати за описание на дигитални обекти за всяка обособена информационна единица, с цел да се улесни индексирането ѝ респективно намирането ѝ в цифровата библиотека. Интерес предизвика и подход за автоматично разпознаване и коригиране на ръкописи на глаголица. Специалистите, участвали в създаването му, са работили при проектирането на метода с документи в много лошо физическо състояние. Съответно, комерсиалните продукти за оптично коригиране и разпознаване са били абсолютно неприложими. А ръчната корекция от специалисти е била много трудоемка и с голям процент на неточност поради избледнялото мастило, петна и увреден физически носител. Приложението предоставя по няколко вероятностни модела на всеки символ от глаголицата, изписван в документите, и съответно „карти на близост“ на всеки един символ. С помощта на моделите и въз основа на матрица с най-голямото съвпадение или приближение, след локализиране на увредени места се възстановява увреден символ.

В сесията за езикова обработка и кодиране (Linguistic Processing & Encoding) бяха представени проекти и експерименти за автоматично извличане на мета данни от предварително оптично разпознати исторически текстове и за автоматичен лингвистичен анализ и нормализация на стари езикови форми на думи от стари документи. Екипите от специалисти са били провокирани от факта, че за много от дигиталните копия на документи, предоставени в интернет, мета данните са изчезнали, а за други те са неточни, непълни и неуеднаквени в различни библиотечни каталози. Изследователите подчертаха, че коректно асоциираните мета данни с дигиталното копие на документ, осигурява по-добър достъп и разкриване на тези колекции. В тази сесия проект представи и екип от български специалисти, което естествено предизвика голяма гордост у мен. Нашите математици и експерти в областта на компютърната лингвистика и езикова обработка, представиха своя метод за корекция и нормализация на думи от стари текстове. Той се основава на автоматично извличане на различни езикови вариации на една и съща дума от различни текстове, декомпозирането ѝ, анализ на структурата ѝ и сравнението ѝ с варианти на думата в съвременния език. Тяхното решение е приложено върху така наречения Шекспиров английски език и с

негова помощ са нормализирани над 50 000 думи от ранно-модерния английски от 17 век с над 80% точност.

В сесията, озаглавена Postcorrection, бяха представени разработки за допълнителна корекция на автоматично разпознати и коригирани текстове. Едно от приложенията предизвика големия интерес на аудиторията, първо защото е разработка на студенти и второ защото те са изнамерили един на пръв поглед много лесен, но хитър и продуктивен начин за корекция на грешките от прилагане на OCR програмите. Чрез софтуера се дава възможност на потребителите, които провеждат търсене в пълнотекстови документи сами да въвеждат предпочитани от тях думи, фрази, с които да се подменят некоригирани или грешно коригирани такива, както и да валидират предложени от други потребители корекции. Методът е разработен като средство прототип чрез стандартни технологии – Java, Ajax. Младите разработчици обещаха неговото публикуване като пакет с отворен код до края на годината с работно заглавие corr4ocr. Още един софтуер за пост корекция беше представен PoCoTo. Тази интерактивна система е разработена в рамките на проект IMPACT и е с отворен код като целта е да се използва и прилага върху различни сканирани и разпознати оптично писмени документи, да се разширява и оптимизира като функционалност. Един от мощните инструменти в PoCoTo са множеството вградени правописни езикови речници, чрез които бързо и лесно се подменят допуснатите от автоматичното разпознаване грешки в документите. Експериментално софтуерът е приложен за пост корекция върху дигиталните масиви на три големи европейски библиотеки. Резултатите са показали, че времето необходимо на специалисти за ръчна корекция на разпознатите тестове, е значително редуцирано. Системата е публикувана като инструмент с отворен код на сайта GitHub.

В сесията, предназначена за споделяне на добри практики и опит (Best Practices and experiences), бе представен и проект, успешен и значим с резултатите си, в който отново участва български специалист – Симона Стоянова, възпитаничка на Катедрата по класическа филология към СУ, професионално реализирала се и участвала в множество проекти към Катедрата по дигитална хуманитаристика в Университета в Лайпциг. Екипът представи работата си над създаване на каталог, съдържащ стандартни мета данни за описание на документи на гръцки и латински. Ценното в този каталог са изградените релации между оригинала на един документ, неговите по-късни издания и преводи на други езици. Каталогът обхваща над 3679 антични гръцки и латински произведения, предоставя мета данни за документите във FRBR формат и създадени по стандарт TEI XML файлове с транскрипции и препратки между различните най-общо казано версии на метаданните на обхванатите документи.

Публикациите от конференцията, представящи проектите и резултатите от реализацията им, са пълнотекстово достъпни от сайта на дигиталната библиотека The ACM Digital Library на Асоциацията за изчислителна техника (Association for Computing Machinery).