Оцифроване надбання: збереження, доступ, репрезентація [1]

Методика оцифрування документів

Одним з характерних трендів у гуманітарній сфері на початку третього тисячоліття стали процеси оцифрування і представлення в електронному середовищі об'єктів історико-культурної спадщини: музейних артефактів, книжкових зібрань, архівних документів. Спочатку це була невід'ємна частина створення електронних каталогів зібрань найбільших бібліотек, музеїв та архівів, електронні копії виконували роль презентаційних (мультимедійних) продуктів, але цифрові проекти досить швидко набули статусу самостійного, самодостатнього, надзвичайно капітало- і наукоємного напряму в діяльності фондоутримувачів. Це пов'язано не стільки з організацією «широкого доступу» до фондів, скільки з необхідністю забезпечити фізичне збереження оригіналів шляхом створення їх електронних копій та вилучення оригіналів із читальних залів, а також виконанням державних планів і програм з перетворення історико-культурного надбання у цифровий формат.

Можна виділити два типи вимог до оцифрування залежно від цілей:

1. Обов'язковий – отримання копій сторінок у вигляді графічних зображень, здійснюване шляхом сканування з подальшою обробкою і збереженням в одному з форматів графічних файлів. У цьому випадку повністю зберігається оригінальна верстка книги і виключається спотворення змісту. Можливість пошуку по тексту відсутня.

2. Пошуковий (факультативний) – оптичне розпізнавання тексту з подальшим збереженням розпізнаного тексту як підкладки набору графічних символів. У цьому випадку стає можливий повнотекстовий пошук по книзі.

Оцифрування документів «своїми силами» ставить перед установою такі завдання:

- визначення параметрів відбору документів для оцифрування;
- аналіз параметрів документів з метою визначення матеріально-технічної бази, необхідної для оцифрування документів та вимог, що пред'являються до електронного контенту;
- забезпечення матеріально-технічної бази (придбання обладнання, його налаштування);

---

[1] 16–17 грудня 2014 р. у Державному політехнічному музеї при НТТУ «КПІ» було проведено другий науково-практичний семінар «Оцифроване надбання: збереження, доступ, репрезентація». Тема заходу – «Оцифрування як інформаційне виробництво і сервіс». Мета заходу – ознайомлення фахівців з передовими світовими практиками, навчання технологіям оцифрування, формування та інтеграції цифрових колекцій, методам їх розповсюдження та використання. Другий семінар присвячений питанням організаційно-технологічного і нормативного забезпечення процесів оцифрування і організації сервісів із цифровим контентом. Посилання: Оцифроване надбання: збереження, доступ, репрезентація http://museum.kpi.ua/conferences/digital-heritage-2014/

● навчання кадрів, що займатимуться оцифруванням;

● введення комплексу оцифрування в експлуатацію, стажування кадрів;

● за необхідності організація додаткових курсів підвищення кваліфікації;

● визначення планових показників, впровадження системи звітності;

● навчання спеціалістів центрів оцифрування у напрямі обробки цифрових ресурсів – каталогізації.

Інший спосіб оцифрувати свої фонди це скористатися послугами організацій, що займаються оцифруванням. Наприклад спеціалізований центр «БАЛІ» (http://scbali.com/) пропонує широкий спектр послуг, серед яких не тільки послуги з оцифрування і обробки зображень, фото послуги, а й дистрибуція, локалізація і супровід спеціалізованого програмного забезпечення для електронних бібліотек, краєзнавчих ресурсів, електронних енциклопедичних довідників; розробка, впровадження та супровід технологічних рішень для електронних бібліотек і цифрових колекцій; послуги з оцифрування, обробки зображень, очищення й розпізнавання текстів; спільні проекти і партнерство у створенні інтегрованого цифрового контенту, удосконалення систем повнотекстового пошуку в багатомовних ресурсах і веб-репрезентації оцифрованих об’єктів.

Через необхідність розробки інструментальних засобів контролю якості оцифрованих зображень та нормативної документації для країн СНГ, ЗАО «ДиМи-Центр» (http://www.dimi.ru/dc/), за підтримки Російської державної бібліотеки (РДБ) та Державної науково-технічної бібліотеки Росії (ДПНТБ Росії) розробили методичні вказівки і технічні інструкції з оцифрування «Методика контроля качества сканирования бумажных документов» (http://www.dimi.ru/dc/index.php?option=com_content&view=article&id=106).

У процесі роботи вивчався закордонний досвід розробки універсальних засобів об’єктивного контролю якості сканування документів. Було враховано результати проектів Metamorfoze (http://www.metamorfoze.nl/, Нідерланди) і The National Digital Information Infrastructure and Preservation Program (http://www.digitalpreservation.gov/, США) та ін.

Як джерело освітлення можна використовувати «холодні» електролюмінесцентні або флуоресцентні лампи з фільтром, що захищає від ультрафіолетового випромінювання і поглинає тепло, також можна використовувати світлодіоди, або волоконно-оптичні системи освітлення. Освітлювачі мають не завдавати шкоди об’єкту сканування потоком ІЧ і УФ випромінювання. Ще одною вимогою до освітлення є забезпечення рівномірності освітлення по всьому формату оригіналу для отримання якісних результатів і збереження природного світло-тіньового балансу. У деяких бібліотеках газети, архівні матеріали, рукописи розшивають перед оцифруванням, а потім не зшивають, зберігають у спеціальних картонних коробках виготовлених з безкислотного картону.

Технічні засоби обробки і управління оцифрованими документами

Компанія ABBYY (http://www.abbyy.ua/), провідний світовий розробник програмного забезпечення та постачальник послуг у галузі лінгвістики, розпізнавання документів та введення даних. Своєю місією компанія обрала допомогу людям розуміти один одного. Створюючи рішення в галузі штучного інтелекту, введення документів, обробки даних та перекладу, ABBYY перетворює інформацію у корисні знання.

ABBYY розробляє ключові технології у чотирьох напрямах:

● розпізнавання документів (OCR – оптичне розпізнавання тексту; ADRT – розпізнавання структури документу; MRC – зменшення розміру зображення при зберіганні у PDF; перетворення PDF-файлів);

● потокове внесення даних (ICR – оптичне розпізнавання символів, написаних від руки; OMR – розпізнавання міток; OBR – розпізнавання одновимірних і двовимірних штрих-кодів; класифікація документів; інтелектуальний аналіз сторінки; вилучення даних з будь-яких типів документів);

● аналіз та розуміння тексту (інтелектуальний корпоративний пошук; вилучення даних; eDiscovery. Перспективні напрями: класифікація документів; порівняння документів; аналіз тональності висловлювань; охорона інформаційного периметра організації з виявленням фактів передачі несанкціонованої інформації; система прогнозування та оповіщення про події; багатомовний пошук);

● лінгвістика та переклад (морфологія, синтаксис, семантика; управління термінологією; впровадження лінгвістичних технологій; послуги з локалізації; переклад документації; лінгвістична підтримка міжнародних заходів; створення глосаріїв і баз пам’яті перекладу; електронні та друковані словники).

Автоматизація введення документів і даних:

● Переведення оцифрованих документів, зображень та PDF-файлів у редаговані формати;

● Створення електронних архівів з можливістю швидкого пошуку потрібних документів;

● Введення даних в інформаційну систему підприємства з будь-яких видів паперових бланків, анкет та фінансових документів.

Лінгвістичні рішення і послуги:

1. ABBYY Language Services (http://abbyy-ls.com.ua/) – Лінгвістична підтримка корпоративних клієнтів:

● локалізація програмного забезпечення, веб-сайтів, мультимедіа і маркетингових матеріалів з англійської більш ніж 80 мовами;

● послуги письмового перекладу великих обсягів документації. Лінгвістичний супровід великих міжнародних заходів, послуги перекладу телефонних переговорів, відео-конференцій.

2. Видавництво ABBYY Press (http://www.abbyypress.ru/) – випуск словників, енциклопедій, а також робота над путівниками.

3. ABBYY Compreno (http://www.abbyy.ru/isearch/compreno/) – Система розуміння, аналізу та перекладу текстів на природних мовах.

Спеціалізований центр «БАЛІ» розробив систему формування та інтеграції цифрових колекцій DC-Visu (http://demo.dcvisu.com/) онлайнових репрезентацій оцифрованих документів. DC-Visu реалізує: репрезентацію оцифрованих документів, якісно відтворюючи його в цифровій формі, яка передає всі особливості зовнішнього вигляду оригіналу; посимвольне подання тексту для повнотекстового пошуку; візуалізацію оригіналу, пошук та інші можливості роботи з текстом в єдиному інтерфейсі; представлення документа єдиним об'єктом, у якому зосереджено кілька електронних форматів; розвантаження користувача: інтуїтивно зрозумілий інтерфейс без локальних інсталяцій; використання наявних систем для бібліотечних сервісів; збагачення інформаційного змісту та візуальних сервісів. Система управління електронними колекціями розрахована на використання в електронних бібліотеках. Може бути використана як інструментарій для перегляду зовнішніх об'єктів у бібліографічних та повнотекстових інформаційно-пошукових системах.

DC-Visu V 2.0 дає змогу керувати колекціями оцифрованих документів, сумісна із, популярною у бібліотеках країн СНД, СУБД «ІРБІС»; надає можливість цитування книг; використовує ABBYY-сервіси для розпізнавання тексту, для покращення пошуку; підтримує можливість імпорту метаописів у форматі XML у колекції Europeana (http://www.europeana.eu/).

У наступних версіях планують реалізувати можливість представлення різних типів документів, наприклад листових. Крім того, слід зазначити, що всі розробки орієнтовані на проект Europeana, тому всі мета описи є і будуть сумісними з вимогами цього проекту ***(Підсумки другого науково-практичного семінару «Оцифроване надбання: збереження, доступ, репрезентація» // Національна бібліотека України ім. В. І. Вернадського (http://www.nbuv.gov.ua/node/1969). – 2015. – 10.02).***