УДК [004.352:004.62]:004.652

Георгій Асєєв,
доктор технічних наук, професор,
завідувач кафедри інформаційних технологій ХДАК

Концепція компонента уведення електронних документів у повнотекстову базу даних

Розглядається один з компонентів створення нетранзакційних електронних сховищ будь-яких повнотекстових або графічних даних — масового уведення паперових документів. Наводяться характеристики відсканованих документів різних класів, вимоги до їхньої підготовки для сканування, уведення та обробляння документів, документа, організації робочого процесу, контролю за виконанням та переведенням документа на зберігання в повнотекстову базу даних.

Ключові слова: *архів, бібліотека, електронний документообіг, електронні сховища, опрацювання документів, сканерні технології, технологія оптичного уведення.*

Раніше автор розглядав різні підходи до архітектури й побудови електронних сховищ даних [1—11]. Ці підходи мали на увазі роботу із транзакційними сховищами: банківськими системами, мережами супермаркетів, великими підприємствами тощо. У цій статті проаналізуємо, як функціонує електронний документообіг при створенні повнотекстових баз даних у бібліотеках і великих архівах різних організацій, що виключають будь-які транзакції.

Як правило, подібний електронний документообіг містить компонент первинної (попередньої) роботи з документами: їхнє масове уведення, реєстрацію, обробляння, управління роботою, контроль за виконанням й переведення документа на зберігання в повнотекстову базу даних. Нині усе більше бібліотек і великих архівів використовують сканерні технології для вирішення завдань управління документообігом, що постають перед ними.

Сканер (scanner) — пристрій для перетворення графічної інформації в електричний сигнал. Стосовно цього напряму — для перетворення в цифрову форму й уведення в комп'ютер різної неструктурованої інформації, а конкретніше — для оцифрування зображень.

Масове уведення паперових документів. Під сканцентром розуміється деякий структурний підрозділ або самостійна організація, що забезпечує реалізацію сканерних технологій, тобто виконання організаційно-технічних заходів щодо перетворення інформації з паперових або інших носіїв у комп'ютерні формати. Вихідними документами при цьому можуть бути різні книжкові тексти, бланки, довідки, анкети тощо. Велика продуктивність сканцентра досягається за рахунок використання сучасних мережевих технологій, що забезпечують створення єдиного інформаційного простору й автоматизацію документопотоку, та заміни ручного уведення технологією оптичного уведення.

Масове уведення паперових документів за допомогою сканерів з метою отримання зображень — складний і трудомісткий процес, що потребує досить значних початкових вкладень, а також постійних поточних витрат на функціонування системи. Використовуючи сучасні технології для підвищення продуктивності роботи операторів і автоматизуючи велику кількість ручних операцій процесу уведення документів, можна значно знизити поточні витрати і зробити виправданим застосування систем масового уведення документів.

Спочатку кілька слів про документи, якими наповнене українське бібліотечне й архівне діловодство різних організацій, з погляду на їх виконання. Вони можуть бути віднесені до одного з наведених далі класів:

- "гладкі" тексти (власне текст без графічних ілюстрацій і таблиць);
- документи зі складною топологією й графічними ілюстраціями (включаючи логотипи та підписи);
- таблиці;
- документи, надруковані на гербовому фоні (цінні папери);
- документи з нестандартним розташуванням полів (стародавні книги, різні документи тощо);
- стандартні форми (банківські, податкові, страхові декларації);
- документи з рукодрукованими (handprinting) символами (символи вписуються від руки у виділених полях по нанесеній пунктиром сітці);
- різноформатні документи (від візитівок до креслень формату A0);
- рукописні документи.

Особливу проблему становить електронне архівування креслень для потреб систем автоматизованого проектування (САПР). За даними International Data Corporation і журналу Document Management, у світі налічується понад 8 млрд креслень, і тільки 15% з них перебуває в CAD-Форматах (тобто придатних для роботи в CAD-Системах) [12—13]. Певна річ, у різних країнах справи йдуть неоднаково — у США, наприклад, у CAD-Форматах, зберігається 45% усіх креслень. Опублікованих даних по нашій країні немає, але можна припустити, що в Україні стан не кращий загальносвітового. Сьогодні майже для всіх очевидна необхідність впровадження електронного документообігу, і безсумнівно, що сканування є найефективнішим засобом перетворення паперового архіву в електронний.

З іншого боку в геоінформаційних системах (ГІС) коло завдань, розв'язуваних із застосуванням сканерів, у порів-

нянні із САПР, набагато ширше. Основні класи картографічних матеріалів з погляду технології сканування такі:

— топографічні карти масштабів від 1:25000 до 1:1000000 на папері або плівці, що використовуються для створення топооснови;

— тематичні карти загального застосування на папері (геологічні, екологічні, економічні та десятки інших);

— планшети масштабів 1:500, 1:2000, 1:5000 і 1:10000 як на гнучких матеріалах (папір і плівка), так і на твердих основах (алюміній, картон, фанера, дерево);

— аеро- і космічні фотознімки на плівці й папері.

До топографічних карт в усьому світі пред'являються найвищі вимоги з точності — загальна погрішність не має перевищувати 0,1 мм. При скануванні аеро- і космічних фотознімків у різних завданнях ставляться жорсткі вимоги щодо точності (від 10 до 50 мкм) і дозволу зображення (від 5 до 25 мкм).

Відповідно до поліграфічного виконання документи можна умовно поділити на документи гарної, середньої й низької якості. Однією із найвужчих ланок технології електронного архівування є сканери, що забезпечують безвідмовне високопродуктивне масове уведення в систему документів, виконаних на паперових носіях низької якості: злиплих, вицвілих, порваних, різних розмірів і щільності, погано надрукованих, забруднених тощо.

Уведення документів у систему поділяється на декілька стадій: підготовка документів до сканування; отримання зображення документа; уведення даних, що містяться в документі; експортування даних і публікація.

Підготовка документів до сканування — важлива фаза процесу уведення в систему документів. Ретельна підготовка документів для сканування забезпечує одержання достовірних відсканованих зображень, що зберігаються в системі, і є ключовим чинником ефективності системи, вона відбувається поетапно: визначення самого документа; вибір конкретних галузей його застосування та визначення технологічного ланцюжка руху. Слід розуміти, що відбувається з документом і коли треба з'ясувати, хто приймає рішення з появою спірних документів, а також накреслює конкретні дії в складних ситуаціях; в якому стані підготовка документів для сканування: відкриття конвертів, видалення скріпок або інших предметів, що заважають процесу; встановити робоче місце для сканування. Потрібно визначити місця для документів, призначених для сканування, і для вже відсканованих документів; підготовити пакети документів для сканування. А перед цим необхідно відсортувати документи різних класів і сформувати їхні пакети. Кожний пакет може супроводжуватися спеціальним титульним аркушем з його ідентифікаційним кодом. Все це дає змогу одночасно сканувати пакети з документами різних класів без додаткових затримок, що сприяє підвищенню ефективності системи, зростанню пропускної здатності та прискорює доставляння документів до робочого місця для сканування.

Процес отримання зображення документа складається зі сканування, обробляння зображень, контролю якості відсканованих зображень і не виключає можливість повторного сканування. Сканування — відповідальна операція, отже, до вибору конкретної моделі сканера необхідно підходити уважно. Варто враховувати такі фактори, як розміри документів, їхній стан, чи є документ однобічним або двостороннім, визначити необхідний дозвіл зображення тощо. І отут потрібна допомога експерта. Нині на ринку пропонується така велика кількість різних моделей сканерів, отож нелегко зорієнтуватися в їхньому розмаїтті. Ціна сканера не має бути єдиним критерієм вибору. Слід пам'ятати доречну російську приказку "Скупой платит дважды". Вибираючи сканер потрібно керуватися такими критеріями:

Можливості сканера при роботі з різними типами та кількістю документів (Paper handling). Необхідно проводити оцінку пристрою з урахуванням можливості сканувати без "зажовування" "реальні" документи, що не завжди бувають в ідеальному стані. Перевагу варто віддавати тому, що може працювати з документами різної якості, розміру й ваги, а також зводить до мінімуму ймовірність їхнього "зажовування". Варто також звертати увагу, наскільки зручно звільняти сканер від "зажованого" паперу.

Продуктивність. Швидкість сканера не єдиний критерій оцінки його продуктивності. Необхідно враховувати, що результуюча продуктивність складається із продуктивності сканера, оператора, який його обслуговує, тому важливо, наскільки комфортно фахівцеві працювати зі сканером і управляти ним, наскільки зручний доступ до вже відсканованих документів тощо. Реальна продуктивність сканування залежить на 80% від швидкості сканера, навіть при добре налагодженому технологічному процесі. Хоча варто сказати, що 50—60% — реалістичніша оцінка результуючої продуктивності.

Якість відсканованих зображень. Більшість сканерів мають додаткові плати для поліпшення її якості. Результати праці значною мірою залежать від типу та стану відсканованих документів, і тому оптимальний варіант — тестування конкретного пристрою на реальних документах.

Надійність сканера складно оцінити, не випробувавши його у вирішенні власного завдання. Не маючи такої можливості, потрібно покластися на досвід консультантів, а також на репутацію фірми-виготівника.

Використання різних методів для поліпшення отриманих зображень: вирівнювання, поворот, застосування різних фільтрів, видалення фону, необхідно з метою:

— *поліпшення читаності зображення.* Оброблені зображення зрозуміліші при візуальному перегляді;

— *підвищення точності розпізнавання.* Застосування спеціальних методів поліпшення зображення може значно підвищити точність оптичного розпізнавання символів (OCR). До цих методів можна віднести відновлення зруйнованих символів, видалення ліній, шуму тощо;

— *зменшення розміру зображення.* Розмір файлів оброблених зображень може бути меншим первісного на 80%. Зменшення розміру зображення розуміється як простий стиск файла, так і видалення непотрібної інформації, як-то: таблиць, затінення, логотипів тощо. Контроль зображень необхідний для того, щоб усі потрібні документи були відсканові й легко читані (не повинно бути пропущених сторінок, неякісних зображень тощо).

Для підвищення ефективності та надійності системи варто мати можливість вибіркової перевірки якості відсканованих зображень, а при скануванні об'ємних документів — можливість відслідковувати порядок сканованих сторінок.

Проблема повторного сканування виникає через незадовільну якість зображення або через проблеми, пов'язані з неправильним порядком сторінок у документі (наприклад, за сторінкою під номером 3 іде сторінка 5). Помилки, пов'язані з порушенням порядку сторінок у документі, найнеприємніші та найтрудомісткіші у виправленні й, як правило, потребується повторне сканування.

Добре спроектовані системи потокового уведення мінімізують необхідність повторного сканування. Правильний вибір сканера й програмного забезпечення — запорука ефективності системи.

Уведення даних, що містяться в документі. Існують три підходи до уведення даних.

Уведення ключових слів. Одне або кілька ключових слів використовуються як індекси для конкретного зображення. У подальшому можливий швидкий доступ до зображення документа із застосуванням уведених ключових слів/індексів.

Уведення всього тексту документа. Виконується уведення усіх слів документа й після цього можливе здійснення повнотекстового пошуку його зображення. Цей метод може застосовуватися у разі необхідності одержання текстового варіанта документа.

Формоорієнтоване уведення даних. Застосовується для повної заміни ручного уведення даних у комп'ютерні системи й, в основному, для уведення даних з форм (стандартних, однотипних документів). У цьому випадку зображення служить засобом одержання даних і надалі може бути спрямоване на архівування або навіть вилучення.

Метод уведення даних визначається вимогами конкретного додатку, але при будь-якому підході завжди включає такі кроки:

1. *Безпосереднє уведення даних.* Найпоширеніший метод уведення даних — по відсканованому зображенню (key from image). Для автоматизації процесу уведення можуть бути використані такі технології розпізнавання: друкованого тексту; окремо розташованих рукописних букв і цифр; штрих кодів; "галочок" тощо.

2. *Перевірка даних*, що вводяться, — необхідний атрибут систем уведення. Проникнення в систему некоректних даних може призвести до сумних результатів. Для важливих документів у системах ручного уведення застосовується метод подвійного уведення для забезпечення більшої надійності даних. У цьому випадку два різних оператори здійснюють уведення одного й того ж документа. Якщо дані збігаються, то вони визнаються правильними, а інакше незбіжне поле визнається некоректним і відправляється на спеціальне обробляння.

Системи автоматичного розпізнавання звичайно разом з результатом повертають так звану "ступінь упевненості". Символи або слова, позначені як правильно прочитані, відправляються до оператора для перевірки й корекції. У цьому випадку він бачить реальне зображення документа й може корегувати неправильно або непевно розпізнані символи. Однак не у всіх випадках системи розпізнавання відзначають їх як некоректні. Названі помилки систем розпізнавання є найбільш критичними.

Основний фактор при оцінці ефективності систем розпізнавання полягає у вартості виправлення помилок при розпізнаванні, а не в точності й швидкості системи. В окремих випадках витрати на виправлення помилок при розпізнаванні можуть перекрити всі плюси автоматизації й зробити ручне уведення по зображенню ефективнішим.

3. *Обробляння форм.* Питання уведення даних з форм вимагає спеціального й детальнішого обговорення. Для автоматизації цього процесу необхідне використання досконаліших технологій розпізнавання й ретельнішого проробляння технології функціонування всієї системи в цілому. Процес розробляння документа є дуже важливим, бо від того, наскільки правильно, з погляду конкретних методів розпізнавання, розроблений документ для сканування, залежить відсоток помилок при розпізнаванні даних. Якість друку відіграє велику роль. Вигідніше витратити трохи більше грошей на друкування документа, а потім заощадити значні суми на скануванні й розпізнаванні. Більш високі вимоги в цьому випадку висуваються до методів перевірки даних, що вводяться. Для підвищення їхньої надійності застосовуються додаткові механізми — словники і таблиці, що обумовлено користувачем. Як правило, системи включають спеціальні вбудовані засоби для визначення необхідних процедур перевірки для кожного поля документа.

4. *Експортування і публікація.* Заключна стадія процесу — експортування зображень документів і супутніх даних у конкретну систему документообігу або базу даних. Головними вимогами до експортування є підтримка різних форматів даних і його швидкість. Потокове уведення має своєю метою надати дані для подальшого використання. Таким чином, на завершальному етапі необхідно перевести інформацію в зручний варіант, щоб забезпечити колективний доступ до неї користувачів і можливість оперувати нею за допомогою прикладних додатків. Необхідно записати образи документів у файлову систему (на сервер, на локальний жорсткий диск або на магнітооптичну бібліотеку), індексні дані — у базу даних системи управління документами або будь-якою іншою базою даних, а розпізнаний текст — у файли різних форматів.

Основні характеристики ефективної системи уведення:

— відкритість. Можливість включити в систему різні технології та програмні продукти залежно від конкретного додатка, навіть якщо ці продукти поставляються іншими фірмами. Необхідна можливість інтеграції з різними Workflow-системами та з системами документообігу;

— можливість настроювання. Інтерфейс користування має настроюватися для досягнення максимальної ефективності роботи операторів;

— масштабованість. Необхідно мати можливість додавати й зменшувати системні ресурси з різними рівнями завантаження системи;

— можливість адміністрування. Користувачу потрібно надати можливість гнучкого управління системою, а також контролювати використовувані ресурси й інструментарій для одержання різних видів звітів;

— захист інформації в системі. Система повинна мати розвинені засоби розмежування доступу користувачів до електронного архіву даних, забезпечувати призначення й розмежування повноважень доступу користувачів до документів архіву відповідно до встановлених прав і регламенту доступу. Потрібні засоби контролю всіх дій (журналізація, трасування, аудит) користувачів при роботі з архівом, включаючи аудит сесій, запитів, відновлення бази даних тощо.

Управління процесом уведення в електронне сховище складається з визначення різних потоків і фаз обробляння документів, призначення цих потоків конкретним операторам, розподілу роботи між операторами.

Технології сканування й типи застосовуваних сканерів ми розглянемо в наступній публікації.

Список використаної літератури

1. *Асєєв Г.* Методологія створення сховищ даних. Проблема виявлення нового знання методами Knowledge discovery in databases / Георгій Асєєв // Вісник Книжкової палати. — 2008. — № 4. — С. 23—26.
2. *Асєєв Г.* Методи добування та знаходження знань у сховищах даних електронного документообігу / Георгій Асєєв // Вісник Книжкової палати. — 2008. — № 9. — С. 29—31.
3. *Асєєв Г.* Концепція електронного сховища даних / Георгій Асєєв // Вісник Книжкової палати. — 2009. — № 2. — С. 28—30.
4. *Асеев Г.* Методы интеллектуального анализа данных в электронных хранилищах / Георгий Асеев // Бионика интеллекта. — 2009. — № 1. — С. 28—33.
5. *Асєєв Г.* Методологія створення сховищ даних. Стандарти та моделювання / Георгій Асєєв // Вісник Книжкової палати. — 2009. — № 5. — С. 30—32.
6. *Асєєв Г.* Основні компоненти інформаційного сховища / Георгій Асєєв // Вісник Книжкової палати. — 2010. — № 2. — С. 30—33.
7. *Асєєв Г.* Архітектура корпоративного сховища даних / Георгій Асєєв // Вісник Книжкової палати. — 2010. — № 10. — С. 20—25.

8. *Асеев Г.* Методы интеллектуальной предобработки данных в электронных хранилищах / Георгий Асеев // Радіоелектроніка. Інформатика. Управління. — 2010. — № 2. — С. 106—111.
9. *Асеев Г.* Методы интеллектуального анализа данных в электронных хранилищах. Генетические алгоритмы / Георгий Асеев // Радіоелектроніка. Інформатика. Управління. — 2011. — № 2. — С. 82—86.
10. *Асєєв Г.* Вітрини даних — необхідна ланка в концепціях побудови сховищ даних / Георгій Асєєв // Вісник Книжкової палати. — 2012. — № 7. — С. 26—29.
11. *Асеев Г.* Нейросетевой анализ непрерывных потоков данных из электронных хранилищ / Георгий Асеев // Системи обробки інформації. — 2013. — Вип. 4. — С. 52—56.
12. *Андронов Г. Д.* Современные методы ведения архивов на базе программно-технических средств ARCIS / Г. Д. Андронов // Делопроизводство. — 1999. — № 1. — С. 27—33.
13. *Медведева Г. А.* Автоматизированные системы учета и поиска документов в Российском госархиве научно-технической документации / Г. А. Медведева // Отечественные архивы. — 2001. — № 3. — С. 73—78.
14. *Носевич В. Л.* Архив электронных документов: белорусский опыт / В. Л. Носевич // Отечественные архивы. — 2002. — № 1. — С. 44—52.

Рассматривается один из компонентов создания нетранзакционных электронных хранилищ каких-либо полнотекстовых или графических данных — массового ввода бумажных документов. Приводятся характеристики сканируемых документов различных классов, требования к их подготовке для сканирования, вводу обработке документов, организации рабочего процесса, контроля за исполнением и переводом документа на хранение в полнотекстовую базу данных.

One of the components of the development of non-transactional electronic warehouses of full text and graphic data is examined, namely mass input of paper documents. Characteristics of scanned documents of various types, requirements for their preparation for scanning, document input, document processing are presented. Operations administration, execution control and transfer of documents to the full text database are described.

Надійшла в редакцію 15 вересня 2013 року