УДК 004.352.:004.62

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

Концепція технології масового уведення документів в електронні архіви

Об'єктами технології масового уведення даних можуть бути друковані (книги, брошури, журнали, збірники статей, листівки, рекламні матеріали тощо), рукописні документи, раніше відсканованні зображення, мультимедійні джерела, а також інші електронні ресурси. Перетворення будь-якого документа в електронний формат може вважатися повноцінним, якщо відповідає певним вимогам: структуроване подання документа і наявність програм оптичного розпізнавання. Представлено аналіз характеристик і об'єктів розпізнавання. Подані рекомендації щодо вибору сканера.

Ключові слова: *електронні системи зберігання інформації, об'єкт сканування, програми розпізнавання, архів.*

У попередній статті[*] було обіцяно присвятити наступні публікації технологіям сканування документів для електронних архівів. Далі подаємо відомості з цієї проблеми.

Одним із найважливіших завдань впровадження систем електронного документообігу (СЕД) є переведення паперових документів в електронний формат. Системи документообігу тісно пов'язані зі створенням електронних архівів, що можуть бути складовими або ж окремими незалежними проектами. Розвиток ринку архівування й потокового уведення документів стимулюється нагромадженням в управлінських системах електронних документів, що найчастіше не мають паперових аналогів. Повсюдне впровадження електронних архівів і систем документообігу збільшує потребу в опрацюванні великого масиву наявних і нових паперових документів. Проблема переведення документних фондів в електронну форму нині надзвичайно актуальна. Основними складовими цього процесу є сканування й розпізнавання електронних образів документів.

Сканування документів, тобто одержання образів паперових документів в електронному вигляді, є невід'ємною частиною процесу переходу від паперового ведення справ і паперових архівів до електронних систем зберігання інформації й роботи з нею. Об'єктами сканування можуть бути як прості посторінкові тексти на аркушах різних форматів (плакати, газети, листівки, буклети та ін.), так і складні, зброшуровані (зшиті) документи (книги, брошури, збірники статей, журнали та ін.), що мають велику кількість реквізитів і зовнішніх оцінок, а також рукописні документи, раніше відсканованні зображення, мультимедіа-джерела й інші електронні ресурси.

Під потоковим (або масовим) уведенням звичайно розуміють наповнення електронних архівів шляхом здійснення сукупності операцій сканування, індексації, створення електронних ресурсів з образів документів різного типу.

Одним із найбільш трудомістких і тривалих етапів впровадження будь-якої системи електронного документообігу є первинне уведення інформації в СЕД, чи сканування й створення мінімальної картки електронного документа з 4—5 атрибутами, чи повне розпізнавання тексту документа.

Масове уведення даних у систему автоматизації документообігу не обмежується додаванням значень у таблиці баз даних і копіюванням нових файлів у сховища. Дані вводяться з урахуванням таких, погоджених і затверджених, компонентів:

- організаційної структури компанії й системи;
- єдиних довідників і класифікаторів, що використовують у системі;
- форматів і типів файлів, затверджених внутрішніми положеннями про діловодство в організації.

Термін "масове уведення даних" має відносно давнє походження. Він виник на зорі впровадження інформаційних систем, коли розпочалось активне використання ємних носіїв інформації (магнітних стрічок, барабанів і дисків). Саме тоді з'явився інтерес до технологій, пов'язаних з автоматичним уведенням інформації. Розроблялися спеціальні зчитувальні автомати, побудовані на логічних схемах, призначених для уведення певних типів символів, зокрема, цифр. Сучасні технології оптичного розпізнавання текстів (англійське скорочення "OCR" — Optical Character Recognition) істотно перевершують можливості перших автоматів. З їх розвитком значно скоротилося застосування праці друкарок в операціях по уведенню великих обсягів даних.

Проблема масового уведення даних стала особливо актуальною останнім часом. Очевидно, що для організації управління й контролю потрібно, щоб будь-яка інформаційна одиниця, що містить дані фізичної особи або підприємства, які потрапляють у зону інтересів держави (фінансова операція, акт реєстрації чого-небудь тощо), відразу з'являлася в інформаційних системах і була доступна в разі потреби. Можливість спільної роботи розподілених по всій Україні інформаційних систем — питання окреме, однак для початку потрібно мати ефективний та маловитратний механізм збору інформації. Дані від підприємств можуть бути відразу представлені в електронному форматі, у той час як більшість громадян подають їх на папері. Інформація на паперових носіях фактично використовується лише в поточному діловодстві. Широке застосування державними й комерційними структурами наявної паперової інформації вимагає системного перетворення її в електронний формат. З огляду на той факт, що обсяг нагромадженої інформації на паперових носіях величезний, найбільш раціональним є залучення сканерів при уведенні даних в електронні системи. Але це тільки один з етапів

[*] *Асєєв Г.* Концепція компонента уведення електронних документів у повнотекстову базу даних / Георгій Асєєв // Вісник Книжкової палати. — 2013. — № 11. — С. 20.

перетворення — будь-який документ, що потрапив в електронний світ, вважається повноцінним лише коли відповідає певним вимогам. А одною з головних вимог є структуроване подання документа, що уможливлює його пошук, сортування й модифікацію. Другим етапом є конвертування документа в текстовий файл за допомогою програм розпізнавання. Щоб успішно побудувати систему уведення інформації, слід чітко уявляти завдання, що потребують вирішення.

Завдання уведення даних

Інформація, яку вводять, буває двох типів: структурована й неструктурована. У першому випадку це анкети, таблиці й інші форми, в яких дані рознесені по полях і типізовані, тобто прописано, який тип даних (прізвище, дата, вид документа тощо) повинен бути в кожному полі. Неструктурована інформація — це звичайний текст.

Зупинимося на завданні уведення структурованих даних (спрощено назвемо його "уведення форм"). Під час цієї операції виникає серйозна проблема — досягнення якості інформації. Інформація є неякісною, якщо дані не відповідають вимогам їх подання або об'єктивним правилам, закладеним у природі даних. Приміром, якщо дата має бути зазначена у вигляді "день—місяць—рік", а заповнена форма "місяць—день—рік", то це приклад помилкового подання даних. Якщо ж у номері місяця написано "13" або не збігаються підсумки, то це вже порушення правил, закладених у смисл даних.

Коли відбувається збір даних в електронному форматі, є можливість контролювати їх на етапі уведення або конвертації. Приміром, оператор може вибрати дані за певний місяць із наявного списку й шляхом перевірки відповідності отриманої суми підсумків уникнути помилок. Коли ж людина заповнює паперовий бланк, а механізми перевірки відсутні, то робить це відповідно до свого розуміння. Тому ми кажемо, що інформація у паперовому форматі має суттєво нижчу якість, ніж в електронному. Звичайно, що кінцева якість інформації істотно залежить від того, як зроблений сам бланк, що має пряме відношення до організації процесу уведення документів.

Сканери

Сканери для масового уведення даних поділяють на дві категорії: швидкісні й промислові. Різниця між ними полягає насамперед в їх продуктивності. У швидкісних сканерів — діапазон від 10 до 40 сторінок у хвилину, у промислових моделей — від 40 до 200 і більше.

Зовні швидкісні сканери схожі на звичайні офісні сканери, обладнані механізмом автоматичної подачі документів, проте продуктивність їх значно вища. Більшість швидкісних сканерів можуть працювати як у режимі автоматичної подачі аркушів, так і в режимі планшетного сканування (їх використовують для сканування книг і журналів). Швидкісні сканери випускають компанії Fujitsu, Bell+Howell, Mitsubishi, Hewlett-Packard, Avision, Kodak та ін. Ці пристрої позиціонуються передусім для офісного застосування в робочих групах. Однак, завдяки їх відносній дешевизні, багато організацій в Україні віддають перевагу саме таким сканерам для обробки великого обсягу документації. Автоподатчики швидкісних сканерів уміщають від 50 до 100 сторінок.

Промислові сканери (в англійській термінології — production scanners) відрізняються від швидкісних тим, що мають значну механічну міцність і можуть працювати в безперервному режимі з максимальною швидкістю сканування до 500 сторінок на хвилину. Автоподатчики дозволяють завантажувати до 500 і більше сторінок. Промислові сканери випускають компанії Kodak, Banctec, Bell+Howell, Fujitsu (модель M3099) та ін.

Зазвичай промислові сканери мають додаткове обладнання, що дозволяє вирішувати спеціалізовані завдання, наприклад, спеціальні лампи для сканування кольорових машиночитанних бланків.

Убудований принтер (imprinter) дає можливість друкувати в куті сканованої сторінки текст, що потім з'явиться на відсканованому зображенні.

Важливі характеристики

Єдина загальна умова для сканера, призначеного для роботи з великими обсягами документів — автоматична подача паперу. Інші залежать від поставлених завдань. У певних випадках форма документів розробляється під сканер, у інших, навпаки, — сканер підбирається під документи (якщо формат не можна змінювати за жодних умов, тоді сканер добирають експериментально).

Якісні сканери коштують недешево, тому купувати їх треба з перспективою на майбутнє. Припустимо, потрібно перевести в електронний формат документи з уже наявного архіву. Як правило, архівні папери зберігаються підшитими й пропустити їх через автоподатчик неможливо. Здавалося б, оптимальним рішенням є придбання кількох планшетних сканерів, які будуть паралельно обробляти архівні документи. Насправді ж має сенс придбати один швидкісний сканер з автоподачею сторінок, що може працювати в планшетному режимі. При уведенні документа із планшета швидкісний сканер працює набагато швидше звичайного, отже, потреба в кількох пристроях відпаде і також буде економія на робочих місцях. А нові, ще не підшиті документи можна сканувати з автоподачею. Отже, придбавши один швидкісний сканер, процес уведення як нових, так і старих документів можна значно оптимізувати.

Слід враховувати, що продуктивність, зазначена в паспортних даних сканера, може бути досягнута тільки в тому випадку, якщо всі інші чинники близькі до ідеальних. Для того, щоб завантажити на повну потужність сканер із продуктивністю 200 сторінок у хвилину, варто мати у своєму розпорядженні бригаду добре навчених робітників, які будуть чітко, без метушні й затримок готувати документи до сканування. При цьому самі документи мають бути однорідними, а папір міцним, рівно обрізаним, без ворсу й рваних країв. На практиці ж швидкість сканування залежить від багатьох чинників, що не мають безпосереднього відношення до сканера, тому його реальна продуктивність визначається тільки в роботі. Приблизні дані про реальну швидкість можна одержати, розділивши паспортну продуктивність сканера навпіл, при цьому слід враховувати ще одну особливість. Річ у тім, що найчастіше паспортна швидкість досягається при скануванні аркушів формату А4 в альбомній орієнтації (тобто за мінімальної поздовжної довжини аркуша) з роздільністю 200 сегментів на дюйм. Між тим, для більшості завдань масового уведення цієї роздільної здатності недостатньо, оскільки губляться важливі деталі зображення. Збільшення ж роздільності до 300 точок на дюйм може призвести до падіння швидкості сканування в півтора-два рази. Крім того, для практичних цілей навряд чи знадобляться зображення в альбомній орієнтації — для їх повороту доведеться використовувати або спеціальні плати (що коштує недешево), або досить потужний комп'ютер, що буде повертати зображення в темпі сканування.

Ресурс роботи сканера визначається, виходячи із загального обсягу документів, які потрібно перевести в електронну форму. Нормальний строк — 5—7 років. Відробивши цей час без значних простоїв, пристрій, незалежно від вартості, окупить себе багаторазово.

Ресурс сканера не завжди зазначено в паспортних даних. Однак цю інформацію можна одержати від дилера. Варто дізнатися про ресурси витратних компонентів сканера. Наприклад, ролики й інші деталі механізму подачі, залежно від моделі сканера, розраховані на сканування від 100 до 500 тис. аркушів. А лампу при інтенсивній

експлуатації, швидше за все, доведеться міняти раз у півтора-два роки, а може й частіше. Також, купуючи сканер, має сенс запастися основними витратними компонентами принаймні на рік.

При виборі сканера важливо оцінити характеристики вихідних документів і вимоги до їх збереження. Приміром, старі документи не витримають проходу через автоподатчик, що вибирає сторінки з пачки за рахунок тертя (а саме таким типом пристроїв оснащена більшість сканерів). Цей автоподатчик також не підходить, якщо скановані документи мають значну цінність і не можна допустити їх ушкодження. У цих випадках використовують сканер з подачею на електростатичній стрічці (до якої аркуш ніби "прилипає") або з вакуумним підсмоктуванням сторінок. Ці сканери коштують досить дорого, але забезпечують максимальне збереження документів. Правда, варто відзначити, що вакуумна подача найчастіше не є автоматичною — аркуші кладуть у приймальний лоток окремо.

Надійність сканера — один із найважливіших параметрів. Це поняття поширюється не тільки на апарат, а й на процес сканування. Іншими словами, збоєм можна вважати не лише несправність сканера (з якісними сканерами таке трапляється вкрай рідко), а й усілякі проблеми в процесі сканування, як-то: "зажовування" паперу, захват кількох сторінок одночасно, перекіс сторінок під час сканування. Ці збої в найкращому разі призводять до істотного уповільнення процесу сканування, а в гіршому — можуть викривити результати. Звичайно, комплекс уведення має бути спроектовано таким чином, щоб максимально нівелювати наслідки збоїв сканера.

Розпізнавання

Ефективність систем розпізнавання при переведенні текстів в електронну форму вже доведена практикою. Однак ще кілька років тому ситуація була протилежною — в ефективність OCR-технологій вірили тільки ентузіасти. Ця ситуація мала цілком об'єктивну причину — наявні на той момент технології були вкрай недосконалими, що не забезпечувало задовільне переведення паперових документів в електронний формат і вимагало ретельної перевірки результатів оптичного розпізнавання.

Перелом у свідомості користувачів відбувся, коли системи оптичного розпізнавання були вдосконалені, а якість електронних документів стала поза критикою.

Об'єкти розпізнавання

Сучасні технології OCR дозволяють досить ефективно розпізнавати друкований текст, незалежно від шрифту, з мінімальною кількістю помилок (наприклад, система FineReader 8.0 і вище).

Якщо проблему введення друкованих текстів можна вважати вирішеною, то опрацювання рукописних текстів — завдання складніше, і ще чекає остаточного вирішення. Уже зараз існують системи, які досить ефективно вводять так званий рукодрукований текст (коли кожна літера в слові пишеться окремо). Ці системи використовуються для уведення даних, написаних від руки. Щоб відрізнити системи, що розпізнають рукописний текст, від звичайних OCR-систем, їх називають ICR (Intelligent Character Recognition).

Прикладом таких даних може послужити анкета застрахованої особи Пенсійного фонду або податкова декларація. Бланки цих форм мають деякі особливості, як-то: виділені області під кожну рукописну букву (знакомісця), реперні чорні квадрати по кутах, чітка інструкція із заповнення — всі ці спеціальні вимоги потрібні для автоматизованого уведення рукописної інформації. Бланки, що відповідають цим вимогам, є машиночитаними.

Завдання уведення неадаптованих рукописних текстів скоріш академічне, ніж практичне. Однак ситуація виглядає зовсім інакше, якщо звернутися до завдання уведення рукописного тексту із планшета або екрана кишенькового комп'ютера. Розвиток ринку кишенькових комп'ютерів багато в чому гальмується відсутністю надійних систем уведення інформації від руки. Почасти це пов'язано з тим, що кишенькові комп'ютери усе ще обмежені в ресурсах, тому повноцінна система розпізнавання на них працювати не може. На думку експертів, протягом найближчих років кишенькові комп'ютери істотно додадуть у характеристиках при збереженні ціни, і тоді можна сподіватися, що й убудовані в них системи розпізнавання рукописного тексту стануть "розумнішими".

Існує великий клас завдань, вирішення яких потребує введення інформації з так званих гнучких форм, повна стандартизація яких неможлива. Типовим прикладом таких форм є банківське платіжне доручення. Інший приклад — обробка вхідної кореспонденції, що вводиться в систему документообігу. У будь-якому листі існують загальні атрибути: відправник, одержувач, дата, номер тощо. Очевидно, що розміщення цих полів у документі може мати багато варіантів, тому для вирішення завдання автоматичного уведення цих атрибутів також потрібна технологія уведення гнучких форм. Вона ґрунтується на описі форми, що містить інформацію, яка може допомогти системі знайти те або інше поле.

Сьогодні у світі немає програмних продуктів, які можна було б настроїти на уведення будь-яких гнучких форм без програмування. Однак уже є готові додатки, підґрунтям яких слугує технологія розпізнавання гнучких форм, така як FineReader Bank — система автоматизованого уведення платіжних доручень. Багато компаній розробляють власні рішення, ліцензуючи у виробників модулі розпізнавання.

Ринок систем потокового уведення щорічно зростає приблизно на 16%, що обумовлено технічними та економічними обставинами. Технічні причини — це інфраструктура комп'ютерів і мереж, що постійно розвивається та дозволяє простіше і швидше передавати розпізнані документи й їх зображення, поліпшити якість сканування при зниженні його вартості, а також підвищити ефективність роботи систем розпізнавання й уведення. Економічною причиною насамперед є потреба суб'єктів господарювання постійно скорочувати витрати на підтримку конкурентоспроможності, у тому числі за рахунок зниження витрат на роботу з паперовими оригіналами документів і змістом паперових архівів організації. Найбільш раціональним у цьому випадку є переведення паперових документів в електронний формат і подальше їх зберігання, що дозволить значно прискорити й спростити роботу з архівом документів.

Объектами технологии массового ввода данных могут быть печатные (книги, брошюры, журналы, сборники статей, листовки, рекламные материалы и т. д.), рукописные документы, ранее отсканированные изображения, мультимедийные источники, а также прочие электронные ресурсы. Преобразование любого документа в электронный формат может считаться полноценным, если соответствует определенным требованиям: структурированная подача документа и наличие программ оптического распознавания. Представлен анализ их характеристик и объектов распознавания. Даны рекомендации относительно выбора сканера.

The objects of the techniques of mass input data can be printed (books, brochures, magazines, collections of articles, flyers, advertising materials, and so on), handwritten documents, previously scanned images, multimedia resources, and other electronic resources. Convert any document into electronic format can be considered adequate if it complies with specific requirements: it is a structured document and present optical character recognition. Presents the analysis of their characteristics and object recognition. Recommendations regarding the selection of the scanner.

Надійшла до редакції 16 квітня 2014 року