УДК 004.056

МЕТОДИКА ОТРИМАННЯ ТАБЛИЧНИХ СТРУКТУР ЗІ СЛАБОСТРУКТУРОВАНИХ ЕЛЕКТРОННИХ ДОКУМЕНТІВ НА WEB-ПОРТАЛАХ ВІДКРИТИХ ДАНИХ

К. т. н. О. А. Блажко, Р. В. Арнаут, М. О. Скрипкін

Одеський національний політехнічний університет
Україна, м. Одеса
blazhko@ieee.org

Запропоновано методику отримання табличних структур зі слабоструктурованих електронних документі,в представлених у форматах DOC(X)/ODT на основі ETL-технології шляхом приведення даних до першої нормальної форми реляційної моделі для їх збереження в CSV-форматі. Методика була випробована при заповненні громадського Web-порталу відкритих даних Одеської області на основі документів DOC-формату та показала скорочення часу процесу заповнення.

Ключові слова: відкриті дані, ETL-процеси, нормалізація баз даних.

У 2015 році з появою Закону України «Про внесення змін до деяких законів України щодо доступу до публічної інформації у формі відкритих даних» Україна приєдналася до всесвітнього процесу структуризації публічної державної інформації [1]. Закон передбачає централізоване розміщення публічної інформації на *Web*-порталах у формі електронних публічних (не таємних, не комерційних, не персональних) однорідних наборів даних, до яких існує *API*-технологія для *Web*/мобільних застосувань.

Національний *Web*-портал створено на основі відкритого програмного продукту *DKAN* за адресою http://data.gov.ua з урахуванням рекомендацій постанови Кабінету Міністрів України [2], основною з яких є надання переваги структурованим текстовим форматам *CSV*, *XML* або *JSON* перед слабо структурованими форматами *DOC(X)*, *XLS(X)* та *PDF*. При завантажені файлів у *CSV*-форматі, який є текстовим еквівалентом реляційної таблиці у першій нормальній формі та може бути створеним із табличних даних документів, дані автоматично стають доступними для *API* у форматах *XML* та *JSON*. Але на поточний момент на порталі із понад 500 наборів даних лише приблизно 50% – це *CSV* та *XML*, що значно зменшує ефективність роботи порталу. Основними причинами цього є велика трудомісткість (кількість часу) ручного процесу перетворення даних з документів офісних систем у *CSV*-формат та наявність помилок користувача, які пов'язані з різними форматами зберігання, типами кодування та структурами таблиць. Відомо, що такий процес перетворення використовує *ETL*-технології, які виконують задачу отримання (*Extract*), перетворення (*Transform*) та завантаження (*Load*) даних [3]. Але програмні рішення *ETL* є складними для звичайних користувачів портала та потребують додаткового налаштування з урахуванням особливостей форматів слабоструктурованих даних, які входять до документів сучасних офісних систем.

Тому метою роботи є скорочення часу на створення файлів *CSV*-формату за рахунок розробки методики отримання табличних структур зі слабоструктурованих документів текстових форматів.

В ході аналізу 90 документів на державних сайтах Одеської області (Департаменти статистики, туризму, охорони здоров'я та екології) було створено класифікацію документів за чотирма класами у порядку зростання складності їх обробки: документ з простою таблицею (20%), документ з багаторядковими назвами таблиць (95%), документ з багатосторінковою таблицею (85%), документ з таблицею, в якій присутні рядки, що характеризують групу рядків з даними (37%).

В результаті проведених експериментів визначено, що процес ручного створення *CSV*-формату містить п'ять етапів: 1) відкриття файлу в текстовому редакторі *Writer* офісного пакета *LibreOffice/OpenOffice*; 2) перенесення знайденої таблиці в редактор електронних таблиць *Calc* (запуск редактора *Calc*, виділення таблиці та копіювання в проміжковий буфер, спеціальна вставка (*HTML*), повторне копіювання при розташуванні таблиць на декількох сторінках); 3) нормалізація

шапки таблиці (об'єднання ієрархії стовпців таблиці, транслітерація назв стовпців, видалення заборонених символів у назвах стовпців, скорочення довжини рядків з назвами стовпців); 4) нормалізація змісту таблиці (перетворення групових рядків в додаткові стовпці, видалення пустих рядків, заміна недопустимих даних з урахуванням типу даних, перетворення форматів дат та чисел з плаваючою комою); 5) збереження електронної таблиці в *CSV*-форматі.

На основі проведених експериментів з ручного перетворення документів різних класів складності було створено методику конвертації регулярних структур даних документів текстових форматів у файл *CSV*-формату, яка дозволила автоматизувати найскладніші кроки ручного перетворення. Додатково методика дозволяє зменшити кількість помилок оператора при ручному перетворенні. З метою уніфікації форматів текстових документів запропоновано перетворювати будь-які текстові документи в єдиний *HTML*-формат, що дозволило створити уніфіковану структурну модель документу, яка охоплює всі проаналізовані елетронні документи. На основі методики розроблено програмне забезпечення з використанням технології *Java* та *Opensource* бібліотеки *LibreOffice/OpenOffice*, що дозволило використовувати методику під будь-якими операційними системами. Програмне забезпечення пройшло апробацію на прикладі десяти документів 1-го класу та десяти документів 4-го класу, витрати часу на перетворення яких представлено в таблиці.

Витрати часу на перетворення документа різних класів складності у різних режимах

Етап	Витрати часу для 4-го класу, хв.		Витрати часу для 1-го класу, хв.	
	Ручний режим	Автомат. режим	Ручний режим	Автомат. режим
1	7	3	5	2
2	18	30	5	0
3	60	5	0	0
4	85	7	0	0
5	5	5	5	5
Всього	175	50	15	7

Експерименти продемонстрували скорочення часу на перетворення документів найскладнішого класу більш ніж у три рази, а для документів найпростішого класу – більш ніж у два рази в порівнянні з ручним перетворенням. В той же час, розроблена методика та відповідне програмне забезпечення має недоліки, які в подальшому необхідно виправити: автоматизація переносу групових рядків в окрему колонку для нормалізації таблиці, розпізнавання багаторядкових назв таблиці, автоматизований пошук продовження багатосторінкової таблиці. Створені набори відкритих даних з файлами *CSV*-формату розміщено на громадському порталі Одеської області за адресою http://data.ngorg.od.ua.

ВИКОРИСТАНІ ДЖЕРЕЛА

1. Про внесення змін до деяких законів України щодо доступу до публічної інформації у формі відкритих даних [Електроний ресурс] : Закон України від 09.04.2015 № 319-VIII. – Режим доступу : http://zakon4.rada.gov.ua/laws/show/319-19

2. Про затвердження Положення про набори даних, які підлягають оприлюдненню у формі відкритих даних" [Електроний ресурс] : Постанова Кабінету Міністрів України від 21.10.2015 № 835. – Режим доступу : http://zakon3.rada.gov.ua/laws/show/835-2015-%D0%BF

3. Марулін, С. Ю. Інформаційна технологія обміну даними між системою електронного документообігу та базою даних інформаційної системи // Дис. канд. техн. наук : 05.13.06 "Інформаційні технології" / Україна, Одес. нац. політехн. ун-т.– 2013.

O. A. Blazhko, R. V. Arnaut, M. O. Skripkin

Method of table structure extracting for semistructured electronic documents on open data web-portal

The authors propose evolution of the ETL-based database matching technology with DOC(X)/ODT formats of text documents and in view of the data to the first normal form of relational model for saving data in CSV-format. The method was tested during filling of the Odessa Region Public Open Data Web-portal based on DOC-format documents and showed reduced time of filling process.

Keywords: *open data, ETL-processes, database normalization.*