УДК 004.624

АЛГОРИТМ ОБРОБКИ ДОКУМЕНТІВ ФОРМАТІВ *DOC(X), XLS(X)* З ТАБЛИЧНИМИ ДАНИМИ НА ОСНОВІ ЇХ XML-ПРЕДСТАВЛЕНЬ

Сіроцінський А.А.
к.т.н., доцент каф. СПЗ Блажко А. А.
Одеський національний політехнічний університет, Україна

АНОТАЦІЯ. Стаття присвячена автоматизації процесів обробки (конвертування) документів форматів *DOC(X), XLS(X)*. Автором запропоновані алгоритми для автоматизованої обробки та уніфікації документів даних різних форматів, що апробовані на прикладах даних з порталу Головного управління статистики Одеської області.

Вступ. Восени 2016 р. Кабінет Міністрів України ухвалив Розпорядження про приєднання України до Міжнародної Хартії відкритих даних, яке передбачає централізоване розміщення публічної інформації на національному *Web*-порталі за адресою *http://data.gov.ua* у текстових форматах *CSV*/*XML*/*JSON*. Але на даний момент лише приблизно п'ята частина документів представлена у вказаних форматах, що пов'язано з великою трудомісткістю (кількістю часу) ручного процесу перетворення даних з документів офісних систем та наявністю помилок користувача через різні формати зберігання, типи кодування та складні структури таблиць. В роботі [1] вже пропонувались алгоритми з обробки документів різних форматів, але результат обробки зберігався в реляційних таблицях, що обмежується використанням лише *CSV*-формату.

Метою роботи є скорочення часу на обробку документів форматів *DOC(X), XLS(X)* на основі їх *XML*-представлень.

Для досягнення мети було поставлено такі **завдання**: створення уніфікованої моделі зберігання документів; створення алгоритмів обробки документів форматів *DOC(X), XLS(X)*; створення алгоритму заповнення уніфікованої моделі зберігання документів.

Опис задач. Для зберігання документів було вибрано *XML*-формат для максимальної універсальності. Для цього було створено теги і побудовано їх структуру в межах клітинки таблиці, її рядка, самої таблиці та всього документа загалом. Отже, було створено наступну структуру зберігання документів:

```
<document>          <!-- початок документа -->
        <tables-number> <Кількість_таблиць>  </tables-number>
        <tables>            <!-- початок таблиць -->
                <table>  <!-- початок таблиці №1 -->
                            <table-title> Назва таблиці 1 </table-title>
                            <table-description>Опис таблиці 1 </table-description>
                            <head-row>        <!-- Початок рядка заголовка таблиці №1 -->
                                    <cell> Вміст і–ї клітинки заголовка </cell>
                            </head-row>       <!-- Кінець рядка заголовка таблиці №1 -->
                            <row>       <!-- Початок рядка даних таблиці №1 -->
                                    <cell> Вміст і–ї клітинки  1–го рядка </cell>
                            </row>      <!-- Кінець рядка даних таблиці №1 -->
                            ……..
                </table>  <!-- Кінець таблиці №1 -->
                <table>   <!-- Початок таблиці №2 -->
                            <table-title> Назва таблиці 2 </table-title>
                            <table-description>Опис таблиці 2 </table-description>
                            ……….
                </table>   <!-- кінець таблиці №2 -->
                ……..
        <tables> <!-- кінець таблиць -->
</document>      <!-- кінець документа -->
```

Рис. 1 – Опис структури *XML*-відображення електроного документу з табличними даними *DOC(X), XLS(X)*

Створення алгоритмів обробки документів. Було створено два алгоритми: окремо для текстових *DOC(X)* і окремо для табличних *XLS(X)* документів.

Алгоритм для обробки текстових документі. Для обробки текстових документів було вибрано формат *ODT*, який має більш наглядну структуру тегів.

Основний алгоритм обробки текстових документів містить наступні кроки.

Крок 1: конвертування документу формату *DOC(X)* в формат *ODT*, засобами *LibreOffice*.

Крок 2: витягання з *ODT*-архіву файлу *context.xls* (дані та основне форматування).

Крок 2: визначення шрифтів заголовків (для подальшого визначення заголовків таблиць).

Крок 4: визначення всіх таблиць документу за відповідними тегами.

Крок 5: запуск циклу по всіх знайдених таблицях: цикл по всіх рядках таблиці, крім «шапки» – витягання даних з клітинок; цикл по рядках «шапки» таблиці – нормалізація даних, запис в *XML*-файл.

Алгоритм для обробки табличних документів. Для обробки табличних документів була використана бібліотека *Apache POI*. Також для скорочення алгоритму було використано *LibreOffice* для конвертування *XLSX* в *XLS*, а сам алгоритм містить наступні кроки.

Крок 1: Якщо документ формату *XLSX*, тоді конвертуємо його в *XLS* засобами *LibreOffice.*

Крок 2: Цикл по всіх листах книги

Крок 2.1: Визначення координат всіх таблиць листа.

Крок 2.2: Цикл по всіх таблицях поточного листа – визначення кількості рядків шапки в таблиці, роз'єднання і заповнення значеннями всіх об'єднаних клітинок таблиць.

Крок 2.3: Цикл по всіх таблицях поточного листа: Запис всіх даних з таблиці в колекцію (крім шапки), визначення шапки таблиці, запис шапки таблиці в початок.

Крок 3: Закриття книги.

Алгоритму заповнення уніфікованої моделі зберігання документів містить наступні кроки.

Крок 1: Створення файлу *XML* за назвою документа, що конвертується.

Крок 2: Якщо файл не існує – створюються елементи (теги) *<document>, <tables-number>, <tables>*, інакше – дописуються дані в існуючий файл.

Крок 3: Створення елемента *<table>*.

Крок 4: Створення елемента *<table-title>* і *<table-description>*, заповнення їх даними.

Крок 5: Цикл по колекції з даними таблиці (рядки таблиці): Створення елементу *<row>* або *<head-row>* (якщо перший рядок), цикл по колекції з даними рядку (стовпці таблиці), створення елементу *<cell>*, заповнення його даними, розміщення елемента в елемент *<row>*.

Крок 6: Зберігання документу *XML*.

Проведення експериментів. Апробація алгоритму проведена на прикладі наборів даних, розміщених на веб-порталі Головного управління статистики Одеської області за адресою http://www.od.ukrstat.gov.ua/. В таблиці 1 розглянуто порівняння затрат часу на обробку наборів даних з вказаного порталу (вручну та за допомогою розробленого програмного продукту).

Таблиця 1 – порівняння затрат часу на обробку даних

№	Назва набору	Кількість таблиць	Час на обробку (вручну)	Час на обробку (програмою)
1	Стан виплати заробітної плати на 1 січня 2017 року	6	16 хвилини	9 секунд
2	Доходи та витрати населення у 2015 році	1	3 хвилини	5,5 секунд
3	Промислове виробництво в Одеській області у 2016 році	2	5 хвилин	7 секунд

ВИСНОВКИ. В роботі було розглянуто алгоритми обробки документів форматів *DOC(X), XLS(X)* для представлення і збереження їх в уніфікованому *XML*-форматі. Експерименти з розробленим програмним забезпеченням показали значне скорочення часу на обробку даних від 33–х до 107–х разів, залежно від розміру документу. В майбутньому планується удосконалити алгоритм для обробки більш рідкісних і нетипових випадків, а також здійснити обробку нових форматів даних *ODT, ODS*.

СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Блажко, А.А. Автоматизация процесса заполнения базы данных на основе электронных документов разных форматов / А. А. Блажко, С. Ю. Марулин, Ю. А. Дунько // Вестник Херсонского национального технического университета. – 2010. – № 2(38). – С. 212-216.