**УДК 004.624**

# ВІЗУАЛІЗАЦІЯ XML-ПРЕДСТАВЛЕННЯ ДОКУМЕНТІВ З ТАБЛИЧНИМИ ДАНИМИ

Янко В. Г.
к.т.н., доцент каф. СПЗ Блажко О. А.
Одеський національний політехнічний університет, УКРАЇНА

**АНОТАЦІЯ.** Стаття присвячена автоматизації процесу візуалізації даних, розміщених на веб-порталах відкритих даних з використанням деревоподібної та табличної структур документу. Автором запропоновано методику автоматичного виявлення помилкових та некоректних даних, апробовану на громадському порталі відкритих даних одеської області.

**Вступ.** Впродовж декількох років залишається актуальною проблема низької якості документів, розміщених на національному веб-порталі відкритих даних за адресою http://data.gov.ua, тому що чиновникам складно виконувати вручну вилучення таблиць з документів та їх перетворення у рекомендовані текстові формати *CSV* або *XML*. В роботі [1] розроблено програмне забезпечення для автоматизованого вилучення таблиць із документів текстового формату, яке зменшило трудомісткість перетворення, але з помилками при наявності складної структури шапки таблиці, декількох таблиць та стилістичних особливостей візуального представлення таблиць користувачем. Тому **метою даної роботи** є зменшення імовірності наявності помилок у документах *CSV*-формату за рахунок розробки формату візуалізації структури вхідного документу та відповідного програмного забезпечення.

**Візуалізація *XML*-представлення документів.** З урахуванням структури вхідних документів у форматах *DOC(X), XLS(X)* (рисунок 1) структуру можна представити в форматі *XML* (рисунок 2), що містить наступні теги: *tables* – кореневий елемент *XML*-структури; *tables-number* – загальна кількість таблиць у документі; *table* – конкретна таблиця з атрибутом згідно номеру таблиці у вхідному документі; *table-title* – заголовок таблиці, *table-description* – опис таблиці; *head-row* – тег, що містить назви колонок таблиці; *cell* – назви колонок таблиці, або дані комірки таблиці; *row* – рядок з даними таблиці.

**Структура зовнішньої торгівлі послугами у січні–червні 2016 року**

| Найменування послуги згідно з КЗЕП | Код послуги згідно з КЗЕП | Експорт | |
|---|---|---|---|
| | | тис.дол. США | у % до січня–червня 2015 |
| **Усього** | | **356944,4** | **81,0** |
| Послуги з переробки матеріальних ресурсів | 01. | 4741,7 | 229,1 |
| Послуги з ремонту та технічного обслуговування, що не віднесені до інших категорій | 02. | 6341,9 | 138,7 |
| Ремонт машин та устаткування | 02.01 | 492,5 | 135,5 |
| Капітальний ремонт транспортних засобів | 02.02 | 3059,3 | 179,2 |
| Поточний ремонт та технічне обслуговування транспортних засобів | 02.03 | 2786,0 | 112,9 |

(а)

```
<?xml version="1.0" encoding="UTF-8"?>
<tables>
  <vizualizer-xml version="1.0_pre"/>
  <tables-number>2</tables-number>
 + <table id="1">
 - <table id="2">
    <table-title>Структура зовнішньої торгівлі послугами
    <table-description>Додаток 2 Структура зовнішньої тор
   - <head-row>
      <cell>Найменування послуги згідно з КЗЕП</cell>
      <cell>Код послуги згідно з КЗЕП</cell>
      <cell>Експорт тис.дол.США</cell>
      <cell>Експорт у % до січня–червня 2015</cell>
      <cell>Імпорт тис.дол. США</cell>
      <cell>Імпорт у % до січня–червня 2015</cell>
      <cell>Сальдо</cell>
    </head-row>
   - <row>
      <cell>Усього</cell>
      <cell>НЕМАЄ ДАНИХ</cell>
      <cell>356944.4</cell>
      <cell>81.0</cell>
```

(б)

Рис. 1 – Приклад вхідного документу (а) та його еквіваленту у розробленому *XML*-форматі (б)

Оскільки чимала кількість помилок містять стилістичні помилки (неправильне об'єднання комірок, розірвані таблиці), програмний модуль, що відповідає за формування розробленого *XML*, працює за рахунок аналізу стилістичного оформлення (жирного шрифту, розриву сторінок) документу у форматі *ODT*, отриманого у результаті конвертації вхідного документу *DOC(X)/XLSX*, а у разі формату *XLS* документ аналізується без конвертації до іншого формату.

Основні етапи роботи розробленої програмної системи наведено на рисунку 2.

Оскільки основною метою візуалізації є полегшення пошуку та виправлення помилок, деревоподібна структура (рисунок 3а) надає змогу швидко оглянути усі наявні у документі дані

та усвідомити, скільки змін необхідно внести. Варто зазначити, що значної кількість помилок користувачеві вдається позбутися в автоматичному режимі без докладання зусиль.

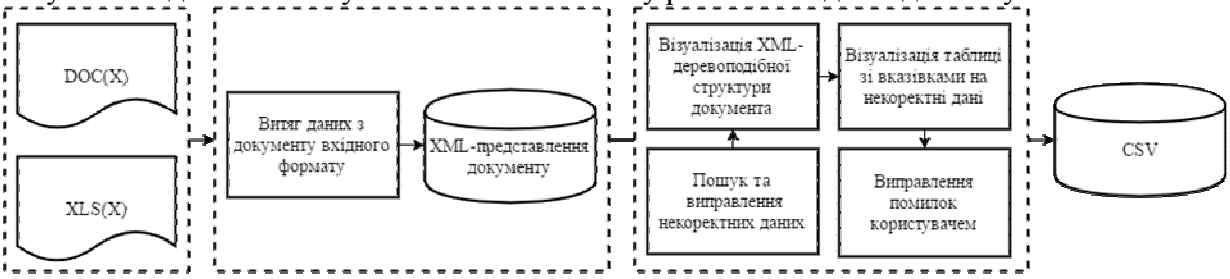
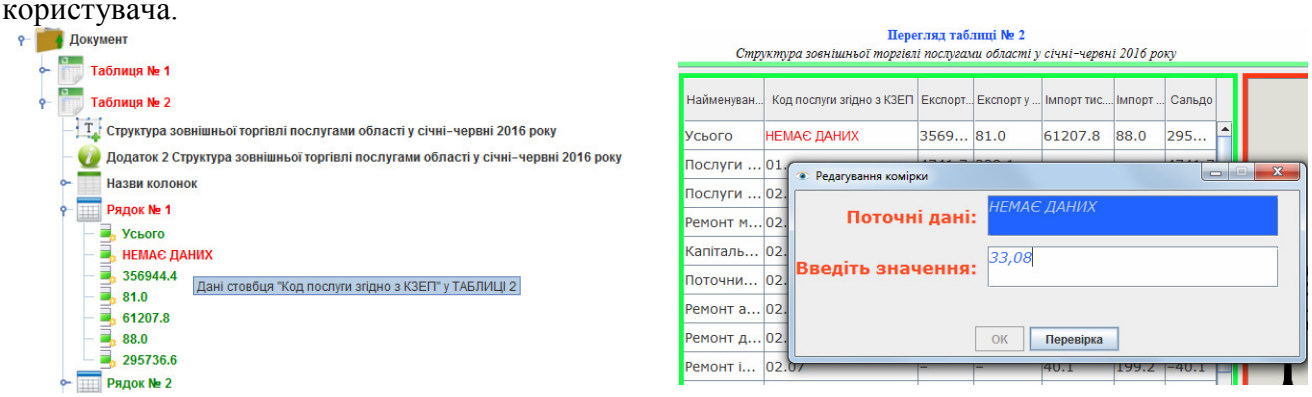



Рис. 2 – Основні етапи роботи програми візуалізації

У разі, якщо автоматично вирішити проблему не вдається, користувачеві надається можливість самостійно виправити дані у табличному режимі (рисунок 3б), який передбачує аналіз внесених користувачем змін та, у разі виникнення помилки, повідомляє про це користувача.




(а) (б)

Рис. 3 – Фрагмент деревоподібної структури документу (а) та приклад режиму редагування (б)

Для апробації роботи автоматизованого процесу візуалізації використано п`ять документів *DOC*-формату, що розміщено на веб-порталі Головного управління статистики Одеської області за адресою http://www.od.ukrstat.gov.ua/. Всі документи містять дві таблиці з ієрархічною структурою заголовку, але різняться кількістю рядків та стовбців в таблицях. Результати автоматизованої обробки документів наведено в таблиці 1, де $E_i$ – кількість помилок при обробці $i$-го документа, $\overline{E}$ - середнє арифметичне кількості помилок. На основі аналізу результатів експерименту можна дійти висновку, що розроблена програмна система зменшує кількість помилок у вхідних документах на 86%.

Таблиця 1 – Аналіз результатів роботи програми

| | $E_1$ | $E_2$ | $E_3$ | $E_4$ | $E_5$ | $\overline{E}$ |
|---|---|---|---|---|---|---|
| Кількість помилок у вхідному документі | 20 | 44 | 16 | 14 | 8 | 51 |
| Кількість помилок після застосування програми | 13 | 18 | 3 | 1 | 0 | 7 |

**Висновки.** У роботі представлено алгоритм роботи програми, що розробляється з метою зменшення кількості помилок у документах, що завантажуються до порталу відкритих даних.

В ході проведення експериментів також з'ясувалося, що існуючі на веб-порталі набори даних містять типові помилки. У майбутньому планується розробити базу знань, що буде містити дані конкретного користувача, на кшталт внесених змін.

## СПИСОК ВИКОРИСТАНИХ ДЖЕРЕЛ

1. Блажко, О.А. Методика отримання табличних структур зі слабоструктурованих електронних документів на web-порталах відкритих даних / О.А. Блажко, Р.В. Арнаут, М.О. Скрипкін // Труды XVII международной научно-практической конференции «Современные информационные и электронные технологии», 23-27 мая 2016 г. – Одеса : Политпериодика, 2016. – С 42-43.