Скрипкін М.О., *магістрант*
Янко В.Г., *магістрант*
Сіроцінський А.А., *бакалавр*
Кафедра системного програмного забезпечення
Одеський національний політехнічний університет

МЕТОДИКА СТВОРЕННЯ НАБОРІВ ВІДКРИТИХ ДАНИХ НА ОСНОВІ СТРУКТУРНО-СТИЛІСТИЧНОГО АНАЛІЗУ ЕЛЕКТРОННИХ ДОКУМЕНТІВ

Запропоновано модифікацію методики та її програмного забезпечення автоматизованого отримання табличних структур зі слабоструктурованих електронних документів текстових форматів офісних пакетів за рахунок підключення стилістичного аналізу тегів ODF-формату. *Апробацію виконано на прикладі національного Web-порталу відкритих даних.*

Ключові слова: *електронні таблиці, відкриті данні, бази даних*

Постановка проблеми та мета роботи. Впродовж останніх років в Україні поетапно впроваджується електронне урядування через успішну реалізацію багатьох *IT*-проектів, одним з яких є створення *E*-сервісів на основі відкритих даних. 22 вересня 2016 р. Кабінет Міністрів України (КМУ) ухвалив Розпорядження про приєднання України до Міжнародної Хартії відкритих даних, що стало юридичним розвитком попередніх рекомендацій з Постанови КМУ [1], які передбачали централізоване розміщення публічної інформації на національному *Web*-порталі за адресою *http://data.gov.ua* у формі електронних публічних (не таємних, не комерційних, не персональних) однорідних наборів відкритих даних (НВД). Основною рекомендацією із постанови було надання переваги структурованим текстовим форматам *CSV*, *XML* або *JSON* перед слабоструктурованими форматами *DOC(X)*, *XLS(X)* та *PDF*, тому що, завантаження документів у перших форматах на *Web*-портал автоматично надає *API*-доступ у форматі *XML/JSON* для *Web*/мобільних застосувань у соціально-економічних сферах, зменшуючи вартість їх супроводу. Але на поточний момент із понад 8-ми тисяч завантажених на портал лише приблизно 17% представлено у необхідному *CSV/XML*-форматі, що підвищує ризик зриву планів реалізації *IT*-проекту створення *E*-сервісів на основі відкритих даних. Основною причиною

цього є велика трудомісткість (кількість часу) ручного процесу перетворення даних з документів офісних систем у *CSV*-формат та наявність помилок користувача через різні форматі зберігання, типи кодування та складні структури таблиць. Рік тому автори цієї роботи створили громадський *Web*-портал Одеської області за адресою *http://data.ngorg.od.ua*, розробили 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 стилистичних особливостей візуального представлення таблиць користувачем [2]. **Тому метою роботи** стало скорочення помилок при створенні файлів *CSV*-формату з табличних структур слабоструктурованих документів текстових форматів на основі структурно-стилістичного аналізу документів.

Результати дослідження. Для досягнення мети роботи запропоновано модифікацію методики автоматизованого отримання табличних структур на основі наступних етапів. Етап 1 – форматна уніфікація документа в *ODF*-формат. Етап 2 – структурний аналіз документа. Етап 3 – стилістичний аналіз документа. Етап 4 – створення деревовидної структури документа в *XML*-форматі. Етап 5 – семантичний аналіз документа з визначення помилок у числах та датах. Етап 6 – візуалізація структури документа з напівавтоматичним виправленням помилок попередніх етапів аналізу. Етап 7 – створення файлу у *CSV*-форматі. На 2-му етапі виконується пошук *XML*-елементів документа *ODF*-формату з урахуванням структурної класифікації тектових документів з табличною структурою, що включає: шапку документа (e1); опис документа (e2); опис таблиці (e3); заголовок таблиці (e4); шапку таблиці, представлену одним рядком (e51) або у вигляді ієрархії (e52); рядок таблиці, що збігається за структурою з колонками шапки (e61) або містить колонки злиття (e62). На 3-му етапі виконується ідентифікація стилів оформлення елементів структурної класифікації з урахуванням стилістичної класифікації текстових документів: шапка документа як простий текст (e11) або як прихована таблиця (e12); заголовок виділено жирним шрифтом (e41) або більшого розміру (e42) у порівнянні з описом таблиці або документа;

таблиця автоматично розташована на декількох сторінках (e81) або з ручним розподілом на сторінки (e82). Апробація методики і програмного обеспеспеченія виконана на прикладі наборів даних національного порталу http://data.gov.ua. Оброблено 7933 наборів даних, які містять посилання на 19082 файли різних форматів, включаючи: *RTF/DOC(X)/ODT* – 33%, *XLS(X)/ODS* – 23%, *PDF* – 15%, *CSV* – 11%, *XML* – 6%. В результаті структурного аналіза документів *RTF/DOC(X)/ODT*-форматів визначено, що 60% з них містять таблиці, із яких 6% – з ієрархічною шапкою. А в результаті стилістічного аналізу визначено, що 16% – приховані таблиці, які можуть стати джерелом помилок.

Висновки. Автоматичне виявлення синтаксичних і стилістичних особливостей вмісту документів дозволило запобігти помилкам при перетворенні документів у файли *CSV*-формату, який є основним форматом подання НВД. В подальшому планується розробити програмне забезпечення етапу візуалізації структури документів та автоматизувати процес розміщення створених НВД на громадському *Web*-порталі Одеської області.

Керівник магістерського дослідження к.т.н., доцент кафедри СПЗ Блажко О.А.

Література

1. 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

2. Блажко, О.А. Методика отримання табличних структур зі слабоструктурованих електронних документів на web-порталах відкритих даних [Текст] / 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.