DOI: 10.37791/2687-0649-2021-16-6-66-83

Программное извлечение данных из word-документов на основе ситуационно-ориентированного подхода

В. В. Миронов[1]***, А. С. Гусаренко***[1]*****, Н. И. Юсупова***[1]

[1] *Уфимский государственный авиационный технический университет, Уфа, Россия*

** gusarenko@ugatu.su*

Аннотация. В статье рассмотрены вопросы применения ситуационно-ориентированного подхода для программной обработки word-документов. Рассматриваемые документы подготавливаются пользователем в среде текстового процессора Microsoft Word или его аналогов и используются в дальнейшем как источники данных. Открытость форматов Office Open XML и Open Document Format позволила применить концепцию виртуальных документов, отображаемых на ZIP-архивы, для программного доступа к XML-компонентам word-документов в ситуационно-ориентированной среде. Обоснована важность выработки предварительных соглашений относительно размещения информации в документе для последующего поиска и извлечения, например, с помощью заранее подготовленных шаблонов-заготовок. Для форматов DOCX и ODT рассмотрено использование ключевых фраз, закладок, элементов управления контентом, пользовательских XML-компонентов для организации извлечения введенных данных. Для каждого варианта построены древовидные модели доступа к извлекаемым данным, а также соответствующие XPath-выражения. Отмечено, что использование того или иного варианта зависит от функциональных возможностей и ограничений текстового процессора и характеризуется различной сложностью разработки шаблона-заготовки, внесения данных пользователем и программирования извлечения данных. Рассмотрен практический пример обработки метаданных научной статьи, подготовленной в среде Microsoft Word для публикации в научном журнале. Примененное решение основано на занесении метаданных в статью с помощью размещенных в шаблоне-заготовке элементов управления контентом, привязанных к элементам пользовательского XML-компонента. Разработанная иерархическая ситуационная модель HSM обеспечивает извлечение XML-компонента, загрузку его в DOM-объект и XSLT-преобразования для получения результирующих данных: отчета об ошибках и JavaScript-кода для последующего использования извлеченных метаданных.

Ключевые слова: ситуационно-ориентированная база данных, иерархическая ситуационная модель, виртуальный документ, открытый текстовый формат, метаданные научной статьи, Open Journal System, DOCX, ODT

Для цитирования: *Миронов В. В., Гусаренко А. С., Юсупова Н. И.* Программное извлечение данных из word-документов на основе ситуационно-ориентированного подхода // Прикладная информатика. 2021. Т. 16. № 6. С. 66–83. DOI: 10.37791/2687-0649-2021-16-6-66-83

Software extract data from word-based documents situationally-oriented approach

V. Mironov[1], ***A. Gusarenko***[1*], ***N. Yusupova***[1]
[1] *Ufa State Aviation Technical University, Ufa, Russia*
** gusarenko@ugatu.su*

Abstract. The article discusses the use of situation-oriented approach to software processing word-documents. The documents under consideration are prepared by the user in the environment of the Microsoft Word processor or its analogs and are used in the future as data sources. The openness of the Office Open XML and Open Document Format made it possible to apply the concept of virtual documents mapped to ZIP archives for programmatic access to XML components of word documents in a situational environment. The importance of developing preliminary agreements regarding the placement of information in the document for subsequent search and retrieval, for example, using pre-prepared templates, is substantiated. For the DOCX and ODT formats, the article discusses the use of key phrases, bookmarks, content controls, custom XML components to organize the extraction of entered data. For each option, tree-like models of access to the extracted data, as well as the corresponding XPath expressions, are built. It is noted that the use of one or another option depends on the functionality and limitations of the word processor and is characterized by varying complexity of developing a blank template, entering data by the user and programming data extraction. The applied solution is based on entering metadata into the article using content controls placed in a stub template and bound to elements of a custom XML component. The developed hierarchical situational model of HSM provides extraction of an XML component, loading it into a DOM object and XSLT transformations to obtain the resulting data: an error report and JavaScript code for subsequent use of the extracted metadata.

Keywords: situationally-oriented database, hierarchical situational models, virtual document, open text format, the metadata of the scientific article, Open Journal System, DOCX, ODT

For citation: Mironov V., Gusarenko A., Yusupova N. Software extract data from word-based documents situationally-oriented approach. *Prikladnaya informatika*=Journal of Applied Informatics, 2021, vol.16, no.6, pp.66-83 (in Russian). DOI: 10.37791/2687-0649-2021-16-6-66-83

Введение

Программное извлечение тех или иных текстовых данных из офисных документов – информационно-технологическая задача, которая встречается достаточно часто. В том числе применительно к документам, созданным в среде текстового процессора Microsoft Word и его аналогов, таких как OpenOffice Writer, LibreOffice Writer, отечественный МойОфис Текст, китайский WPS Office и другие. Подобные документы будем называть word-документами. В этой статье нас интересуют word-документы, которые имеют «двойное назначение»: во-первых, исходный красиво оформленный текст полиграфического качества, во-вторых, источник текстовых данных для дальнейшей автоматизированной обработки. В качестве характерного примера можно привести процесс публикации статей в научном журнале. Авторы оформляют в электронном виде статью в соответствии с установленными требованиями и загружают ее на сайт журнала. Там статья проходит процедуру рецензирования, доработки, исправления, доводится до окончательного вида публикации. Кроме того, в ходе редакционного процесса авторы или редакторы вносят в базу данных многочисленные метаданные статьи, такие как заглавие, сведения об авторах, аннотацию, ключевые слова, спи-