за заг. ред. Л. О. Драгомірової та К. Є. Новохатського. – К.: ДКАУ; УНДІАСД, 2004. – 228 с.

[15] Архіви України [Електронний ресурс]: офіц. веб-портал Державної архівної служби України. – Режим доступу: http://www.archives.gov.ua. – Назва з екрана.

[16] *Алєксєєнко А. О.* Створення веб-сайтів державних архівних установ України: досвід ЦДНТА України // Збереження культурної спадщини в інформаційному суспільстві. Слово архівістів: матеріали Круглого столу, присвяченого 40-річчю ЦДНТА України. – Х.: СПДФО Михайлов Г. Г., 2010. – С. 6–8.

[17] Центральний державний науково-технічний архів України [Електронний ресурс]: офіц. веб-сайт ЦДНТА України. – Режим доступу: http://www.archive.gov.ua. – Назва з екрана.

[18] *Дождьова О. Є.* Документальні виставки on-line як форма використання документної інформації: досвід діяльності Центрального державного науково-технічного архіву України // Від XIX до XXI ст.: трансформація бібліотек у контексті розвитку суспільства: до 125-річчя ХДНБ ім. В. Г. Короленка. – Х.: РА «ІРІС», 2011. – С. 328–330.

В статье рассмотрены основные принципы публикации электронных образов архивных документов в сети Интернет, определены особенности подготовки онлайновых документальных выставок.

Ключевые слова: интернетизация, электронная копия документа, он-лайновая выставка, ЦГНТА Украины.

The basic principles of publishing images of archival documents on the Internet, especially the preparation determined on-line documentary exhibits

Key words: connectedness, an electronic copy of the document on-line exhibition, CSSTA of Ukraine.

УДК [005.92:004.63]:004.932

Андрей Баранцев

ПОИСК И ОТОБРАЖЕНИЕ ИНФОРМАЦИИ НА РАСТРОВЫХ ИЗОБРАЖЕНИЯХ ТЕКСТОВ, СОДЕРЖАЩИХСЯ В ЭЛЕКТРОННЫХ ДОКУМЕНТАХ

Предложен подход к созданию графического редактора растровых изображений, способного осуществлять поиск и выделение текстов на растровых изображениях документов.

Ключевые слова: растровое изображение, электронный архив, текстовая карта.

Непрерывное возрастание объемов архивных документов остро ставит вопрос предоставления оперативного доступа к ним. Наиболее эффективным решением этой проблемы является перевод документов с бумажными носителями в электронный вид и создание на их основе архивов электронных документов. Перевод документа с бумажным носителем в электронный вид подразумевает его сканирование и назначение ему обязательных атрибутов электронного документа. В результате такого перевода появится электронный документ, содержащий растровое изображение. Поскольку такое изображение электронного документа содержит текст, то при работе с ним было бы удобно иметь такой же набор возможностей, как и при работе с обычным текстовым документом.

Ведущие мировые проекты по созданию и накоплению электронных ресурсов (Google Book Search[1], проект Национальной библиотеки Франции Gallica)[2] предоставляют возможности контекстного поиска фрагментов текста и наглядного отображения его результатов на растровом изображении. Это свидетельствует о востребованности подобного рода возможностей. Однако характерной особенностью растровых изображений является необходимость использования сложных процедур для поиска и выделения необходимой информации, а проекты, успешно реализовавшие такие процедуры, не предоставляют описания их реализации.

Целью нашей статьи является ознакомление пользователей и разработчиков электронных архивов с разработанным нами механизмом отображения результатов поиска по содержимому документа на его растровом изображении.

Растровые изображения – это изображения, сформированные из сетки точек, имеющих разную яркость и цвет. Растровая графика позволяет создать (воспроизвести) практически любой рисунок, вне зависимости от сложности, в отличие, например, от векторной графики. Растровые изображения документов создаются, как правило, путем сканирования документа с бумажным носителем – создания его электронного образа с помощью специальных средств, обеспечивающих преобразование изображения на бумажном или ином материальном носителе в цифровую

© *Андрей Баранцев, 2012*

форму. Растровое графическое изображение пригодно для просмотра образа документа с экрана, для распечатки этого образа с сохранением расположения графических элементов, а также для возможного копирования последних в целях создания иных документов. Растровое графическое изображение без этапа распознавания непригодно для какой-либо автоматизированной обработки содержательной части электронного документа.

Результатом распознавания растрового графического изображения является файл, содержащий текст. При этом возможны потери данных о формате шрифта, расположении графических элементов, потеря графических элементов (таких как печать, штамп, герб, рисунок, диаграмма и т. п.). Распознанный текстовый документ частично пригоден для автоматизированной обработки текста. Для того, чтобы распознанный текстовый документ стал полностью пригоден для этого, его необходимо формализовать.

Для задачи поиска текста и его отображения на нераспознанном растровом изображении необходимо знать координаты и размеры всех букв на изображении, а также, каким словам эти буквы принадлежат. Слова на изображении могут располагаться одной частью или разбиваться на несколько частей. Варианты расположения слов на изображении приведены на рис. 1.

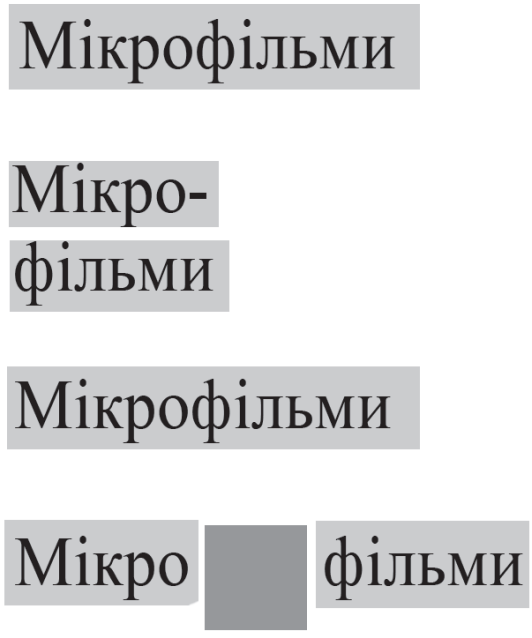

Рис. 1. Варианты расположения слова «Мікрофільм» на растровом изображении текста.

После анализа всех возможных сочетаний вариантов расположения слов на изображении можно выделить формальные атрибуты распознанного документа. Последний состоит из одного и более слов. Каждое слово состоит из одной и более частей, каждая часть – из одной и более букв. Каждая буква занимает на изображении область, которая задается координатами X1, Y1, X2, Y2, где:

– X1 – крайняя левая координата;

– Y1 – крайняя верхняя координата;

– X2 – крайняя правая координата;

– Y2 – крайняя нижняя координата области, занимаемой буквой.

Наиболее подходящим форматом для записи формализованного распознанного документа является формат XML[3]. Преимущества XML – распространенность, открытость и удобство при обработке программным обеспечением. Например, слово «Мікрофільми» после записи в формате XML может принять вид:

```
<ecomap>
<w>
<s>
<l r=«37,123,50,145»>М</l>
<l r=«50,123,62,145»>і</l>
<l r=«65,123,75,145»>к</l>
<l r=«77,123,92,145»>р</l>
<l r=«92,123,97,145»>о</l>
<l r=«97,123,109,145»>ф</l>
<l r=«112,123,124,145»>і</l>
<l r=«127,123,138,145»>л</l>
<l r=«139,123,153,145»>ь</l>
<l r=«155,123,167,145»>м</l>
<l r=«169,123,183,145»>и</l>
</s>
</w>
</ecomap>,
```

где тег «<l>» соответствует каждой букве слова, а ее координаты на растровом изображении описываются атрибутом «<r>» тега «<l>». Тег «<s>» соответствует части слова (в данном случае слово состоит из одной части). Тег «<w>» соответствует слову. Тег «<ecomap>» – устанавливает границы документа.

Распознанный документ, формализованный по указанным выше правилам и сохраненный в файл формата XML, предлагаем называть текстовой картой документа. С ее помощью можно определить координаты области растрового изображения, которую занимает определенное слово или его часть.

Для выделения текста на растровом изображении необходимо сделать копию исходного

изображения и отредактировать его, используя текстовую карту. Редактирование изображения сводится к заливке областей, соответствующих найденным словам, символам и т. п., определенным цветом. Для сохранения оригинальных изображений самих символов необходимо отличать символ от фона и заливать определенным цветом только фон.

Визуальное восприятие символа возможно только при наличии контраста между символом и фоном. Контраст определяется отношением яркости символа и фона. На изображениях текстовых документов яркость фона выше яркости символа. Поэтому при редактировании изображения необходимо выбрать определенный порог яркости и заливать цветом только те точки растра, яркость которых выше порога. Порог яркости должен устанавливаться индивидуально для каждого документа и содержаться в текстовой карте в виде атрибута тега «<ecomap>».

Для отображения результатов поиска по содержимому документа на его растровом изображении нами был разработан неинтерактивный графический редактор-маркер изображения E-Colorer (далее – редактор E-Colorer), демонстрационная веб-страница которого доступна по адресу http://micrography.gov.ua/ecodemo. Основную часть этой страницы занимает растровое изображение документа в формате JPEG[4], на котором демонстрируются возможности редактора E-Colorer. Последний также поддерживает графический формат изображения PNG[5]. В левой части страницы содержится поле для ввода текста, поиск и выделение которого будет производиться нажатием кнопки «Применить». Также в левой части демонстрационной страницы расположены элементы управления, с помощью которых демонстрируются дополнительные возможности редактора E-Colorer. Доступ к этому редактору организован через CGI-сценарий[6], параметрами которого являются:

– имя файла растрового изображения;

– имя файла текстовой карты;

– слово, которое необходимо выделить на изображении;

– ряд дополнительных параметров.

При нажатии кнопки «Применить» редактор E-Colorer загружает текстовую карту и строит по ней объектную модель растрового изображения. Для построения объектной модели используется файл формата XML. Еще может использоваться бинарный файл формата, специально разработанного для хранения тестовых карт, с целью ускорения построения объектной модели.

Объектная модель представления данных оперирует понятиями «класс» и «объект». Классы определяют структуру данных и представляют собой набор атрибутов, которые в свою очередь также могут быть классами, образуя, таким образом, иерархическую структуру. Структура классов объектной модели растрового документа, используемая в редакторе E-Colorer, представлена на рис. 2.

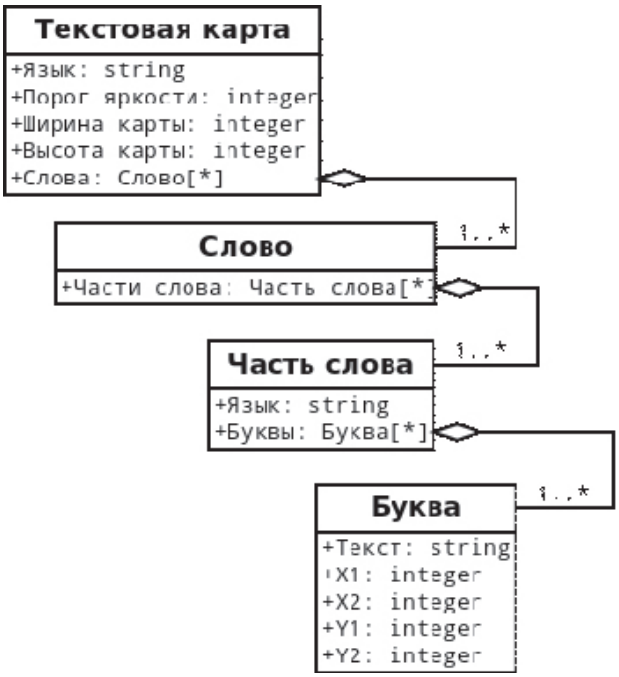

Рис. 2. Структура классов объектной модели растрового документа, используемая в редакторе E-Colorer.

Класс не хранит в себе реальных данных – такую информацию содержат объекты (экземпляры класса). Объектная модель растрового изображения, построенная редактором E-Colorer, состоит из объекта класса «Текстовая карта», который содержит массив атрибутов-объектов, класс которых – «слово». Объект класса «слово» содержит массив атрибутов-объектов класса «часть слова», каждый из которых в свою очередь содержит массив атрибутов-объектов класса «буква». Атрибуты каждой буквы – её координаты на растровом изображении.

Редактор E-Colorer использует объектную модель для поиска заданного слова и вычисления координат всех областей растрового изображения, соответствующих искомому слову. Затем этот редактор делает копию исходного растрового изображения и маркирует вычисленные области цветом, указанным в качестве цвета маркера. Отредактированное таким образом изображение он возвращает CGI-сценарию.

Сейчас редактор E-Colorer реализован в виде CGI-сценария и консольного приложения. Однако внутренняя структура этого редактора такова, что взаимодействие с ним легко реализовать в том виде, в каком это потребуется программе управления электронным архивом. Например, в виде динамически подключаемой библиотеки или веб-службы.

Эффективная работа редактора E-Colorer возможна только при наличии простого способа создания текстовых карт растровых изображений. Благодаря использованию формата XML построить текстовую карту можно и вручную с помощью любого текстового редактора. Координаты символов при этом можно определять с помощью любого растрового графического редактора. Однако это сложный и трудоемкий процесс. Для его автоматизации нами создан редактор текстовых карт E-Mapedit (далее – редактор E-Mapedit). Демонстрационная версия этого редактора доступна для загрузки с FTP[7] сервера по адресу ftp://micrography.gov.ua/pub/demo/ecolorer. Для установки этой демонстрационной версии на рабочий компьютер необходимо воспользоваться инструкцией howto-install.pdf, которая находиться по указанному выше адресу.

Для создания текстовой карты с помощью редактора E-Mapedit необходимо в соответствующем пункте меню указать файл растрового графического изображения в формате JPEG или PNG. Редактор E-Mapedit запускает внешнюю программу распознавания текста на указанном изображении. Для распознавания текста используется программа tesseract[8], которая распространяется под лицензией, допускающей свободное использование программного обеспечения. После завершения распознавания редактор E-Mapedit строит текстовую карту. Результат построения загружается в окно редактирования, в котором оператор имеет возможность, при необходимости, откорректировать текстовую карту и исправить ошибки распознавания, которые могут иметь место в случае плохой читаемости текста на графическом изображении. При сохранении текстовой карты редактор E-Mapedit предоставляет возможность выбора формата файла текстовой карты – XML либо бинарный формат.

Подход, предложенный к формализации атрибутов распознанного текста растровых изображений, открыл возможность реализовать механизм поиска и отображения информации на растровых изображениях текстов, содержащихся в электронных документах. Основными достоинствами реализованного механизма являются:

– простота в использовании;

– нетребовательность к вычислительным ресурсам операционной системы;

– использование открытого формата XML.

Кроме использования реализованного механизма для улучшения качества обеспечения документами пользователей электронных архивов, перспективными также видятся возможности его использования для создания электронных библиотек и в большинстве направлений формирования и использования страхового фонда документации Украины.

[1] Google книги [Электронный ресурс]. – Режим доступа: http://books.google.com. – Загл. с экрана.

[2] Gallica digital library [Electronic resource]. – Mode of access: http://gallica.bnf.fr. – Title from screen.

[3] От англ. eXtensible Markup Language – расширяемый язык разметки.

[4] От англ. Joint Photographic Experts Group, по названию организации-разработчика – один из растровых графических форматов.

[5] От англ. Portable Network Graphics – один из растровых графических форматов.

[6] От англ. Common Gateway Interface – («общий интерфейс шлюза») – стандарт интерфейса, используемого для связи внешней программы с веб-сервером.

[7] От англ. File Transfer Protocol – («протокол передачи файлов») – стандартный протокол, предназначенный для передачи файлов по Интернету.

[8] tesseract-ocr [Electronic resource]. – Mode of access: http://code.google.com/p/tesseract-ocr. – Title from screen.

Запропоновано підхід до створення графічного редактора растрових зображень, здатного здійснювати пошук та виділення текстів у растрових зображеннях документів.

Ключові слова: растрове зображення, електронний архів, текстова мапа.

It is proposed hike to the creation of graphical bitmap editor, capable of search and selection of texts on the raster images of documents.

Key words: raster image, electronic archive, textual map.