УДК: 519.688

Название: Система создания электронных архивов газет с поиском по ключевым словам

Авторы: Марчук А.Г. (Институт систем информатики им. А.П. Ершова СО РАН),
Лештаев С.В. (Институт систем информатики им. А.П. Ершова СО РАН).

Аннотация: В данной статье рассматривается вопрос сохранения архивов газет в цифровой форме. Предлагается технология, охватывающая цикл сканирование-подготовка-публикация, причем в качестве ключевых задач представлены: отображение на сайте материала выпусков газет и поиск статей по заданию текстового поискового образа (ключевых слов).

Ключевые слова: газета, сканирование, публикация, Silverlight, Deep Zoom.

1. Введение. Тенденция создания электронных библиотек на основе цифровых технологий охватывает всё большие предметные области. Данная тенденция актуальна в отношении архивов газет. Важно сохранить архивы газетных подшивок так, чтобы пользователи могли получить доступ к их содержимому современными средствами. Это положение касается не только газетного материала известных изданий, но и малотиражных, ведомственных, заводских и вузовских газет, настенных публикаций. Современные технологии позволяют сохранить образ газетного выпуска и предоставить удобные средства просмотра и поиска по множеству выпусков. Архивы газет, как элементы общей архивной системы, могут быть обработаны для упорядочивания и поиска содержащейся в них информации.

В Институте систем информатики осуществляется проект по формированию современно устроенного электронного архива выпускавшегося с 1961 года еженедельника Сибирского отделения РАН «Наука в Сибири» (до 1983 г. «За науку в Сибири»). С конца 1997 года газета выкладывается на веб-сайте 1. В рамках проекта предыдущие выпуски были оцифрованы, сформирован электронный массив, размещенный на платформе Электронного фотоархива СО РАН http://soran1957.ru .

В статье рассматриваются созданные и апробированные основные технические решения по созданию электронных архивов газет. Достаточно подробное изложение ряда программистских решений приведено в магистерской диссертации [1].

1.а. Научная новизна проекта. В рамках междисциплинарного взаимодействия – информатики и источниковедения осуществляется проект подготовки электронной версии архива еженедельника СО РАН «Наука в Сибири». Проект обеспечит доступность контента в Сети широкому кругу исследователей-историков науки, журналистики, всем

интересующимся историей СО РАН. Для этого используется технология отображения изображений высокого разрешения Deep Zoom.

2. Обзор некоторых известных решений. Корпорация Google разработала своё специализированное решение по публикации выпусков газет – Google Newspapers [5]. В сервисе Google News опубликованы некоторые архивы газет с 1738г. по 2009г [10]. Пользователю предоставляется просмотр сканов выпусков с помощью технологии Google, специально разработанной для этих целей на JavaScript в HTML. За основу взята технология для просмотра карт: изображение разделено на клетки 256x256 разного масштаба, область обзора перемещается с помощью курсора. Добавлена «мини-карта», отображающая положение и размер области обзора относительно изображения выпуска, составленного из горизонтального ряда страниц. Сделана попытка определения позиции и размера заголовков, чтобы при клике на заголовок область обзора перемещалась и масштабировалась на начало статьи.

Сайт issuu.com [12] позволяет публиковать выпуски журналов с помощью технологии, использующей для просмотра Adobe Flash приложение Issuu Reader. Недостатки этого подхода заключаются в следующем: 1) Adobe Flash-приложение при работе нагружает процессор; 2) приложение нельзя модифицировать. Несмотря на это, многие сайты публикуют журналы и комиксы в этой технологии, поскольку она обладает рядом преимуществ. Текст с изображений страниц распознается вместе с позициями слов, что позволяет подсвечивать релевантные слова на изображении при поиске. Источником данных в этой технологии является коллекция страниц выпуска в PDF файле. Интерфейс приложения Issuu Reader позволяет просматривать страницы: коллекцией в виде ровного вертикального ряда, подобно приложениям для просмотра PDF; и по одной, с приближением и анимацией перелистывания страницы.

3. Ввод газетных выпусков и первичная обработка. В созданной в ИСИ технологии обработке подвергается как множество отдельных выпусков газет, так и каждый выпуск в отдельности. Выпуск (номер) газеты в системе структуризации рассматривается как отдельный документ, состоящий из упорядоченного множества страниц. Страницы газеты сканируются и их электронные образы (далее сканы) размещаются в хранилище для последующей визуализации. Кроме того, формируется база данных выпусков газет и их страниц. Выпуск, рассматриваемый как документ, содержит: дату выпуска, название и некоторые другие характеристики, экстрагируемые или заполняемые при сканировании. Обычно название документа совпадает с порядковым номером выпуска в газете. В свою очередь, выпуск рассматривается как документ, состоящий из многих частей, каждая часть

соответствует сканированному участку. Задача упорядочивания частей для получения нумерованной последовательности страниц, возложена на оператора сканирования.

Множество страниц сканируются, изображения подвергаются графической обработке и сохраняются. Преобразование производится в несколько естественных стадий:

1) Листы сканируются с качеством, сохраняющим существенные детали. Выбор параметров сканирования устанавливается заранее. Существенными являются: цветность, разрешающая способность, формат сохранения.

2) Каждый скан листа (разворота) разделяется на две страницы, если сканировался разворот. Иногда такое разделение нецелесообразно, тогда на скане могут остаться две страницы. Страницы обрезаются по краям, упорядочиваются по порядку номеров страниц, при необходимости поворачиваются.

3) Полученные изображения страниц сохраняется в репозитории документов с соблюдением иерархии «газета - выпуск – страница».

Полученные директории со сканами газет превращаются в электронный архив документов и хранятся как исходный материал. Создание и ведение архива выпусков газет, организовано в виде серверного приложения.

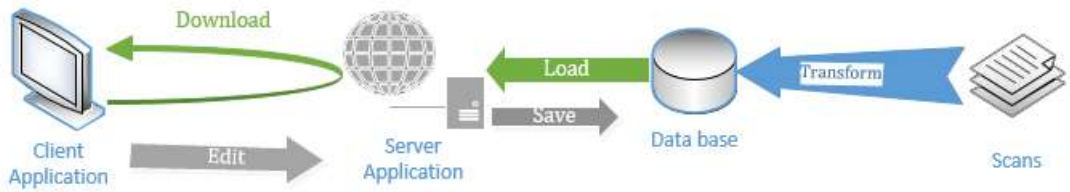

Рисунок 1. Схема web-публикации сканов газет.

На Рисунок 1 приведена общая схема клиент-серверной публикации, в которой электронный архив сканов предварительно преобразуется (Transform) в данные базы данных (Data base). Серверное приложение (Server Application) на запросы от приложений пользователей (Client Application): «получить для просмотра указанный выпуск газеты», загрузит (Load) его из базы данных и отправит в ответ.

Первая стадия схемы публикации – это преобразование электронного архива сканированных страниц в формат базы данных. В качестве базы данных и базы документов (репозитория), используется специальное решение, названное «кассеты» [2]. Такой подход дал возможность распространить способы, технологии и интерфейсы, общие для архивной платформы [3].

Следующие преобразования сканов связаны с проблемами, возникающими на стадии скачивания (Download) приложениями пользователей с серверного приложения. Сканированное изображение в хорошем качестве занимает от 40мб в TIFF формате.

Скачивание пользователем даже одной страницы в таком формате, может занять значительное время. Кроме того, просмотр изображения высокого разрешения в файле больших размеров требует достаточной вычислительной мощности. Нельзя полагаться на то, что конечный компьютер пользователя будет обладать достаточной производительностью и скоростью интернета. Для решения подобной проблемы файл изображения преобразуют с потерей качества в форматы, доступные для демонстрации из браузера: PDF, JPG, PNG. В случае публикации газеты с неудачно подобранными параметрами сжатия, потеря качества изображения для страницы приводит к тому, что шрифт становится сложно читаемым.

Сохранить качество и минимизировать объем передаваемой и обрабатываемой на стороне клиента информации позволяет технология Deep Zoom [11] – решение для web-публикации изображений высокого разрешения от Microsoft. Изображение предварительно преобразуется в формат DZI (*Deep Zoom Image)*, при этом оно разделяется сеткой на кусочки размером 256x256 и каждый сохраняется в формате для быстрого скачивания через интернет JPG или PNG. Имеется несколько вариантов использования технологии для браузеров. Мы использовали технологию Silverlight [17], которая реализована для всех популярных браузеров, не требует от пользовательского компьютера больших вычислительных ресурсов и памяти. Silverlight-приложение в браузере пользователя позволяет обозревать изображение в целом, приблизить некоторую область, при этом приложение скачивает с сервера только попавшие в область фрагменты, по качеству соответствующие масштабу.

Deep Zoom позволяет так же просматривать коллекцию изображений. Для этого используется формат DZC (*Deep Zoom Collection)* – множество расположенных на одной плоскости изображений формата DZI различного масштаба.

Создание DZI и DZC возможно бесплатным пользовательским приложением *Deep Zoom Composer*, кроме того из коллекции изображений пользователь может указать относительные размеры и положения изображений, и получить:

(1) DZC;

*(2)*DZI каждого изображения в коллекции;

(3) DZI всей коллекции, как будто это одно изображение.

При преобразовании архива газет обработка каждого выпуска - коллекции страниц пользователем при создании DZC занимает много времени. Необходим автоматический метод. Deep Zoom Composer предоставляет в SDK .Net библиотеку, с помощью которой в эксперименте было разработано приложение автоматически создающее deepZoomImage и deepZoomCollection. Изображения в коллекции, созданной таким образом, располагаются в горизонтальный ряд без масштабирования.

Изображение в формате DZI можно показывать в Silverlight в компоненте MultiScaleImage или в ASP.NET с помощью компонента *SeaDragon* [16] в наборе AJAX Control Toolkit. При демонстрации deepZoomCollection в Silverlight, можно динамически менять положение и размер каждого изображения в коллекции. Так для демонстрации выпуска страницы выстраиваются в ровный горизонтальный ряд, каждая страница масштабируется до одинаковой константной высоты.

При обработке газеты «Наука в Сибири» выяснилось, что количество страниц архива получается довольно большое, при этом, каждая из страниц, разбивается на множество клеток размера 256x256. При прямом хранении имиджей в файлах, их количество приводит к техническим трудностям хранения и перемещения из-за большого количества файлов (в нашем случае – миллионы). Для решения указанных проблем множество составляющих клеток архивируется в один файл, когда приоритетным фактором является скорость разархивирования. Выбран формат архива без сжатия так, чтобы получить максимальную скорость выборки конкретного изображения. Архивирование произведено по выпускам, соответственно, когда пользователь запрашивает какой-то выпуск, все файлы данного выпуска экстрагируются из архива и сохраняются в кэше.

4. Колонка. Структурно газета рассматривается не только как множество страниц, но и как множество статей. Статья может размещаться на страницах частями: в конце каждой части, не являющейся последней, может быть указана ссылка на продолжение, например, «продолжение в следующем номере»; в каждой части, не являющейся первой, может быть ссылка на предыдущую часть. Следовательно, статья — это упорядоченное множество частей, каждая часть статьи содержит множество колонок, множество изображений с подписями. Кроме того, статья как объект данных имеет идентификатор, как документ имеет заголовок, авторов и дату выпуска первой части. Каждая часть напечатана в некотором выпуске на некоторой странице в некоторой позиции, может начинаться на одной странице, а заканчиваться на другой.

Статья структурно определяется как упорядоченное множество частей статьи в выпусках, причем для каждой части фиксируется позиция и размер охватывающей прямоугольной области.

В интерфейсе Silverlight-приложения прямоугольные области выделения расположены поверх MultiScaleImage в прозрачном контейнере в соответствующих позициях. При перемещении и масштабировании области обзора коллекции изображений в MultiScaleImage положения и размеры областей отражения перемещаются и масштабируются соответственно.

Статьи, персоны, события и организации, их отражения, вносят в данные, указывают положение и размер на страницах газет и редактируют пользователи с правами редактора специальными интерфейсами в приложении клиента.

5. Поиск. Задача поиска информации в тексте газет разделена на подзадачи:

1. Распознавание текста, т.е. отображение из множества сканов во множество текстовых файлов.

2. Индексирование текста поисковым движком с полнотекстовым индексом.

3. Полнотекстовый поиск.

5.1. Распознавание текста. В распоряжении программистов имеется множество программ для распознавания текста (OCR) по изображениям. Анализировались на предмет возможности использования OCR-приложения, удовлетворяющие условиям:

1) Осуществлять распознавание текстов на русском языке;
2) Предлагать бесплатное SDK – комплекта средств разработки инструментов и библиотек. При этом SDK может быть предоставлено сторонним приложением.
3) Предоставлять возможность распознавать директорию со всеми поддиректориями и файлами изображений TIFF.

Были протестированы существующие отдельно от приложений бесплатные движки распознавания текста: Cunei [7] и Tesseract [13]. Имеется также множество приложений и сервисов, которые используют готовые движки. С одинаковыми движками приложения и сервисы распознают текст одинаково.

В специальной литературе имеется ряд публикаций, в которых сравниваются OCR-приложения [18]. На практике качество распознавания зависит от качества печати и сканирования, шрифта и цвета изображения. Поэтому следует с осторожностью полагаться на результаты сравнения в данных статьях, если не известны исходные условия сканирования. Необходимо проверить тестовый образец изображения на множестве подходящих к условиям приложениях. В проведённых испытаниях с использованием приложения Cunei Forms было осуществлено распознавание на тестовом образце изображения большего процента слов, чем с помощью других. По этому показателю движок Cunei был выбран для использования в описываемой системе.

Для программируемого использования Cunei движка в C# имеется библиотека Puma.Net. В числе особенностей Puma.Net, которые оказались полезны:

- позволяет запускать только один процесс распознавания одного изображения, но можно перезапускать приложение для каждой страницы;
- нельзя сохранить результат в DOC формате, но можно в RTF.

- ~1% изображений страниц приводят к ошибкам, из-за которых приостанавливается автоматический процесс распознавания. Решение проблемы – ограничение по времени (до 10 минут) распознавания одного изображения.

Для сохранения распознанного текста файла был выбран формат RTF, поскольку:

- он содержит информацию об относительном расположении текста;
- Существуют .Net библиотеки для извлечения текста из RTF фалов;
- RTF фалы с распознанным текстом имеют приемлемые размеры.

При распознавании текста распознаются фотографии и рисунки на изображении, но для задачи поиска их наличие в выходном файле не требуется.

5.2. Индексирование. *Полнотекстовый индекс (Full Text Index)* – обработка массива текстов таким образом, чтобы, в дальнейшем, эффективно решалась задача лингвистического поиска слова в тексте или поиска по набору ключевых слов. Технически, индексируется текст в столбце таблицы в базе данных. Для наших целей требовались уже готовые базы данных, предоставляющие возможность формирования полнотекстового индекса, поддерживающие русский язык и предоставляющие полнотекстовый поиск. Например, Solr [6], MS SQL 2012 Server Express, и другие [9]. Последняя была выбрана для эксперимента.

Текст располагается в одном столбце блоками: по блоку в строке таблицы. Полнотекстовый индекс применяется к текстовому столбцу, размечая все слова, кроме стоп-слов, которые не несут смысловой нагрузки. Текст предварительно обрабатывается: слова приводятся к нормальной форме.

Для решения задачи поиска в тексте газет устанавливается соответствие слова странице, что позволяет найти множество страниц, содержащих указанное слово.

5.3. Полнотекстовый поиск. К индексированному текстовому столбцу применяются функции *полнотекстового поиска* [8] – по множеству слов определяются строки с ненулевым количеством использований слов в тексте колонки строки. Функции полнотекстового поиска встроены в MS SQL Server 2012 express.

6. Экспериментальное апробирование метода. В рамках проекта «Электронный фотоархив СО РАН» был проведён следующий эксперимент. Были сформированы сканы выпусков(~1700) газеты «За науку в Сибири» с 1965 по 1997. В TIFF формате сканы занимают объем более половины терабайта. Эти же имиджи, переведенные в Deep Zoom формат, составили более 70 гигабайт в JPG формате. На машине с Intel® Pentium®4 2,02ггц. 1,5гб. ОЗУ распознавание заняло несколько недель в результате приостановок из-за ошибок, на индексирование было затрачено несколько часов. Была сформирована база данных, выстроен полнотекстовый индекс.

Запрос осуществляется через задание набора ключевых слов, которые являются поисковым образом. Поиск релевантных запросу страниц занимает несколько секунд. Для данных была использована СУБД, где в RDF формате они сохраняются в XML файлах.

За НАУКУ в СИБИРИ

Орган парткома, комитета ВЛКСМ, объединенного комитета профсоюза, Президиума Сибирского отделения АН СССР.

№ 1 (1). | 4 июля 1961 г., вторник. | Цена 2 коп.

ГАЗЕТЕ «ЗА НАУКУ В СИБИРИ»

Президиум Сибирского отделения Академии наук СССР, партийный комитет, комитет комсомола и объединенный комитет профсоюза горячо приветствуют рождение нашей газеты «За науку в Сибири».

Издание многотиражной газеты — большое событие в жизни Сибирского отделения. Газета нужна всем и от нее многого ждут.

Она должна шире освещать творчество ученых, их упорную работу по координации научных исследований и внедрению достижений науки в производство. В газете необходимо особое внимание уделить выполнению обязательств, взятых учеными в честь XXII съезда КПСС, пропаганде опыта передовых людей науки. Нужно смело и принципиально критиковать недостатки в организации научной работы.

Сибирское отделение не только молодое, но и молодежное. Подавляющая часть наших ученых — молодежь. Воспитывать молодые кадры, освещать их смелые искания, стремление к новым достижениям, к новым открытиям — почетная задача газеты.

Пусть газета борется за улучшение быта, за интересный отдых ученых.

Пусть наша газета верно хранит славные ленинские традиции печати!

Президиум Сибирского отделения АН СССР.

Партийный комитет СО АН СССР.
Комитет ВЛКСМ.
Объединенный комитет профсоюза.

Счастливого пути к читателям

Редакция «Вечернего Новосибирска» сердечно поздравляет многотиражную газету «За науку в Сибири» с выходом в свет ее первого номера.

Восьмитысячный отряд ученых, лаборантов, рабочих и слушащих Сибирского отделения Академии наук получил замечательный подарок — свою газе-

вседневная связь с массами позволяет сделать газету «За науку в Сибири» содержательной, живой, злободневной и острой, подлинным помощником партийного комитета.

Пусть редколлегию нашей газеты никогда не покидает состояние творческого беспокойства, чувство

Президент АН СССР Келдыш Мстислав Всеволодович

Келдыш Мстислав Всеволодович

Рисунок 2. Изображение первой страницы первого выпуска газеты “За науку в Сибири” 1961 г. Выделена фотография президента АН СССР Келдыш Мстислава Всеволодовича. Под выделенной фотографией размещена ссылка на страницу с его информационным портретом.

ГАЗЕТЕ «ЗА НАУКУ В СИБИРИ»

Президиум Сибирского отделения Академии наук СССР, партийный комитет, комитет комсомола и объединенный комитет профсоюза горячо приветствуют рождение нашей газеты «За науку в Сибири».

Издание многотиражной газеты — большое событие в жизни Сибирского отделения. Газета нужна всем и от нее многого ждут.

Она должна шире освещать творчество ученых, их упорную работу по координации научных исследований и внедрению достижений науки в производство. В газете необходимо особое внимание уделить выполнению обязательств, взятых учеными в честь XXII съезда КПСС, пропаганде опыта передовых людей науки. Нужно смело и принципиально критиковать недостатки в организации научной работы.

Сибирское отделение не только молодое, но и молодежное. Подавляющая часть наших ученых — молодежь. Воспитывать молодые кадры, освещать их смелые искания, стремление к новым достижениям, к новым открытиям — почетная задача газеты.

Пусть газета борется за улучшение быта, за интересный отдых ученых.

Пусть наша газета верно хранит славные ленинские традиции печати!

Президиум Сибирского отделения АН СССР.

Партийный комитет СО АН СССР.
Комитет ВЛКСМ.
Объединенный комитет профсоюза.

Рисунок 3. Изображение области Рисунок 2 при приближении на неё с помощью Deep Zoom. Заметно улучшение качества картинки: шрифт менее размыт, его можно прочитать. Приведённое приближение не предельное.

По ключевым словам поиск находит только множество страниц. Пользователю необходимо самостоятельно искать внутри каждой найденной страницы ключевые слова и их контекст. Исследование продолжается в направлении определения и указания их позиций на странице.

Silverlight поддерживается многими браузерами, но только в Microsoft Windows. Например, распространены устройства с touch screen (что подходит для навигации в Deep Zoom) и с операционной системой Android. Для других операционных систем заявлена только альтернатива: Moonlight [14] (от разработчиков Mono [15]). Но текущая версия Moonlight последняя, работа над ней остановлена [4]. Вероятно, и Silverlight тоже последняя версия, поскольку активно предлагается поддержка браузерами стандарта HTML 5, обеспечивающего браузерам большую часть функциональности Silverlight.

Для операционных систем и браузеров, не поддерживаемых для публикации газет, необходимо найти альтернативы Silverlight. Решений может быть несколько: либо «подстроить» систему под Sea Dragon, либо подождать, пока разработчики не начнут поддерживать Deep Zoom Collection. Еще один путь – самостоятельно разработать Deep Zoom альтернативу на HTML 5, который бы поддерживался новыми версиями браузеров.

Заключение. Таким образом, при создании электронного архива газеты «Наука в Сибири» была апробирована технология отображения изображений высокого разрешения Deep Zoom. Использованное решение может быть заменено на Seadragon. Функция поиска по

ключевым словам может быть дополнено определением и отображением позиции на странице, альтернативными программами распознавания текста и индексирования.

Авторы выражают благодарность за помощь в проведении данного исследования сотрудникам Института систем информатики им. А.П. Ершова СО РАН: Фурсенко Алексею Александровичу, Павловской Ирине Юрьевне, Крайневой Ирине Александровне, Филиппову Владимиру Эдуардовичу.

Список Литературы

1. **Лештаев С.В.** Архитектура и программное обеспечение архивных фактографических систем: работа с многостраничными растровыми изображениями: дис… маг. 5.13.11. — Новосибирск, 2012.

2. **Марчук А.Г., Марчук П.А.** Особенности построения цифровых библиотек со связанным контентом // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Сб.трудов / XII Всеросс. научн. Конф. RCDL'2010, Казань, Россия 13–17 октября 2010 г. — Казань: Казан. ун-т, 2010. — С. 19–23.

3. **Марчук А.Г., Марчук П.А.** Платформа реализации электронных архивов данных и документов // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: Труды XIV Всероссийской научной конференции RCDL’2012. Переславль-Залесский, Россия, 15-18 октября 2012 г. – г. Переславль-Залесский: изд-во «Университет города Переславля», 2012, С. 332-338.

4. Мигель де Иказа: “Мы прекратили работу над Moonlight” [В Интернете]. - http://www.linux.org.ru/forum/talks/7810987

5. About Google News Archive Search [Online]. - HYPERLINK "http://support.google.com/news/bin/answer.py?hl=en&answer=1638638&topic=9312&ctx=topic" http://support.google.com/news/bin/answer.py?hl=en&answer=1638638&topic=9312&ctx=topic

6. Apache Solr [Online]. - HYPERLINK "http://lucene.apache.org/solr/" http://lucene.apache.org/solr/ .

7. CuneiForm [Online]. - HYPERLINK "http://en.wikipedia.org/wiki/CuneiForm_(software)" http://en.wikipedia.org/wiki/CuneiForm_(software)

8. Fulltext search [Online] // Wikipedia. - HYPERLINK "http://en.wikipedia.org/wiki/Full_text_search" http://en.wikipedia.org/wiki/Full_text_search .

9. Fulltext search engines [Online] // - HYPERLINK "http://www.mediawiki.org/wiki/Fulltext_search_engines" http://www.mediawiki.org/wiki/Fulltext_search_engines .

10. Google Newspapers [Online]. - HYPERLINK "http://news.google.com/newspapers" http://news.google.com/newspapers .

11. Inside Deep Zoom Part II: Mathematical Analysis [Online] / auth. Gasienica Daniel. - HYPERLINK "http://www.gasi.ch/blog/inside-deep-zoom-2/" http://www.gasi.ch/blog/inside-deep-zoom-2/ .

12. Issuu [Online]. - HYPERLINK "http://issuu.com/" http://issuu.com/ .

13. tesseract-ocr [Online]. - HYPERLINK "http://code.google.com/p/tesseract-ocr/" http://code.google.com/p/tesseract-ocr/.

14. Moonlight [Online]. - HYPERLINK "http://en.wikipedia.org/wiki/Moonlight_(runtime)" http://en.wikipedia.org/wiki/Moonlight_(runtime).

15. Mono [Online]. - HYPERLINK "http://www.mono-project.com/About" http://www.mono-project.com/About.

16. SeaDragon [Online]. // Wikipedia - HYPERLINK "http://en.wikipedia.org/wiki/Seadragon_Software" http://en.wikipedia.org/wiki/Seadragon_Software.

17. Silverlight [Online]. - HYPERLINK "http://www.microsoft.com/rus/silverlight/overview/default.aspx" http://www.microsoft.com/rus/silverlight/overview/default.aspx.

18. Кривошей А. Системы оптического распознавания текста в Linux - обзор и сравнительное тестирование. 2011 [Online] - HYPERLINK "http://rus-linux.net/nlib.php?name=/MyLDP/office/OCR/OCR_review.html" http://rus-linux.net/nlib.php?name=/MyLDP/office/OCR/OCR_review.html

УДК: 519.688

Title: System for creating digital archives of newspapers with keyword search

Authors: Alexander G. Marchuk (A.P. Ershov Institute of Informatics Systems)
Sergey V. Leshtaev (A.P. Ershov Institute of Informatics Systems).

Abstract: This article examines the issue of storage of digital archives of newspapers. Technology is proposed covering scanning-preparation-publication cycle, and as the key challenges presented: presenting online of editions of newspapers and search for articles by keywords.

Keywords: newspaper, scan, publish, Silverlight, Deep Zoom.