

Филиппович Анна

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ ПОДГОТОВКИ И ОБРАБОТКИ ДАННЫХ ДЛЯ ГИПЕРГРАФИЧЕСКОЙ ИНФОРМАЦИОННОЙ СИСТЕМЫ САР

Факсимильная копия страниц САР

Одной из компонент Интегрированной инструментальной информационно-программной среды для автоматизации исследований Словаря Академии Российской является Гиперграфическая информационная система САР. Она содержит набор факсимильных копий изображений страниц Словаря.

Ввод и обработка изображений страниц

Исходными данными для формирования Гиперграфической информационной системы САР (факсимильной копии страниц САР) были ксерокопии страниц словаря. Данные страницы использовались для ввода и редактуры текста словаря, поэтому они содержат метки правки. Некоторые ксероксы страниц очень плохого качества и требуют обработки. Для формирования требований к изображениям и создания технологии ввода и обработки страниц первоначально был проведен пробный эксперимент. Для этого произвольно были взяты несколько ксерокопий страниц разных томов словаря (I, II, IV). Далее страницы были отсканированы на сканере Epson Perfection 2400 Photo. Для сканирования использовались специальная утилита Epson Smart Panel и программа Epson Twain. Используя настройки программы, при многократном сканировании были выявлены оптимальные характеристики сканирования и сформированы требования к результирующим изображениям, а также технология их ввода и обработки.

Далее были отсканированы страницы I, II, III, IV томов САР. В результате полученные изображения содержали ряд недостатков. Главным недостатком изображений была слишком высокая контрастность. Изображения были сохранены в режиме Grayscale, однако выглядели как битовые (Bitmap). Изображения в режиме Bitmap представляются одним цветом, поэтому все пиксели окрашены либо в белый, либо в черный цвет. Изображения в режиме Grayscale (градации серого называются 8 битовыми) и пиксели могут иметь 256 оттенков серого. Это значительно ухудшало визуальные качества изображений: границы букв

имели ступенчатую форму. Исходя из этого, необходимо было обработать изображения, чтобы исправить недостатки оригинала и погрешности сканирования. С этой целью была разработана технология обработки изображений страниц словаря.

Требования к результирующим изображениям

1. Все страницы должны иметь одинаковый размер. Для 1, 2 и 3-го томов: 17см × 21см, для 4 тома: 17см × 22см.
2. Каждая страница должна быть контрастна. Фон — белый, текст черный.
3. Изображения страниц должны быть записаны в режиме Grayscale (градации серого).
4. Разрешение изображений 300 dpi.
5. Изображение каждой страницы записывается в формате *.JPG высокого качества (размер изображения 700-1300 Кб).
6. Все страницы не должны иметь перекосов. Текстовые строки должны располагаться горизонтально, а столбцы вертикально. Допустимая погрешность 0,2°.
7. Все страницы должны удовлетворять требованиям чистоты.

Технология ввода изображений страниц

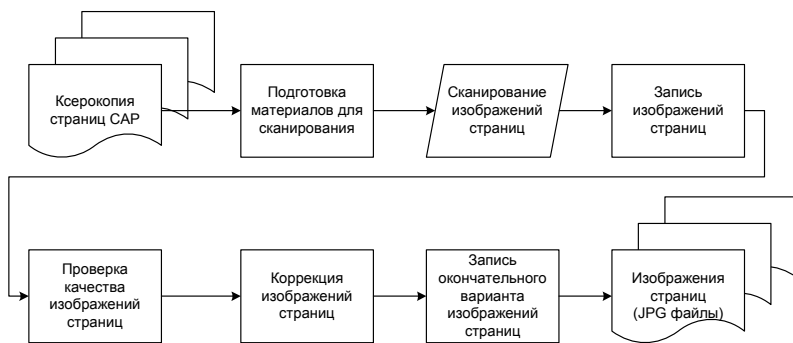


Рис. 1. Технология ввода изображений страниц САР.

Необходимо было отсканировать страницы Словаря Академии Российской, осуществить первичную обработку изображений и записать полученные изображения. В результате выполнения задания по каждому из томов Словаря Академии Российской должен быть сформирован диск с изображениями страниц.

Этапы

1. Подготовка материалов для сканирования:

- Проверка полноты полученных материалов (наличие всех страниц по нумерации).
- Удаление меток правки, выполненных карандашом.

2. Запуск программы сканирования изображений.

3. Сканирование каждой страницы словаря.

Для этого необходимо было выполнить следующие действия:

- Поместить страницу под крышку сканера.
- Осуществить предварительное сканирование.
- Выделить область сканирования так, чтобы был захвачен весь текст.
- Размер сканированной области: для 1, 2 и 3-го томов: 17см (ширина) × 21см (высота), для 4 тома: 17см (ширина) × 22см (высота). Если весь текст не помещается в требуемый размер, необходимо осуществить масштабирование.
- Определить настройки сканирования: тип изображения: черно-белый документ; разрешение: 300 dpi; параметры изображения (экспозиции, гаммы, светлых тонов, темных тонов). Параметры изображения должны были установлены таким образом, чтобы полученное изображение было контрастным. Рекомендованные значения для светлых тонов в пределах 180-250, для темных тонов 40-60.
- Сохранить настройки сканирования.
- Осуществить сканирование.

4. Запись полученных изображений страниц.

Каждая страница словаря должна быть записана в формате *.jpg высокого качества в отдельный файл. Имя формируется следующим образом:

Для страниц словаря: YXXXX.jpg, где Y – номер тома, XXXX – номер первой колонки страницы словаря.

Пример: 60581.jpg – это страница 6 тома словаря, колонка 581.

Для вступительной части (предисловие, изъяснение, краткое начертание, члены академии, показание): YrXXX.jpg, где Y – номер тома, XXX – номер страницы по порядку.

Пример: 6r005.jpg – это 5-ая страница 6 тома словаря.

Технология обработки изображений страниц

Необходимо было обработать отсканированные страницы Словаря

Академии Российской, настроить контрастность, почистить. Обработка каждого изображения включает следующие этапы:

1. Запуск программы Adobe Photoshop.

2. Проверка полученного изображения страницы (страниц) словаря на перекосы.

Текстовые строки должны располагаться горизонтально, а столбцы вертикально. Для того, чтобы проверить существует ли перекося достаточно провести направляющие: вертикальные и горизонтальные.

Если полученное изображение перекошено необходимо его исправить (повернуть). Для этого можно использовать команды Adobe Photoshop:

Image (Изображение) → Rotate Canvas (Повернуть холст) → Arbitrary (произвольно) или Edit (Редактирование) → Free Transform (Свободная трансформация).

Перекосы (искажения) изображений страницы связаны с тем, что при ксерокопировании она лежала неровно. Искажения могли иметь разную форму, но чаще всего трапеции. Некоторые искажения связаны с состоянием страниц оригинала словаря. В таблице 1 представлены примеры искажений страниц.



А) параллельное искажения	Б) трапецивидное искажения
<p>колокольнѣ подѣ голось подобран- ныхъ, одинѣ другаго меньше. Устой церкви огенѣ хорощѣ звонѣ. Звонѣцѣ, нѣдѣ с. м. Сл. Колокольчикѣ, побрякушка. Иѣ окружа его шилки златилии звонции многилии окрестѣ. Сирах. хлв. 11. Звонкѣй, кѣя, кое. Звонкѣѣ, кѣ, ко, прил. Яркѣй, громкѣй звукѣ издающѣй. Коль скоро звонкѣй услышатѣ гласѣ трубы. В. Петр. Ен. Звонко. нар. Звучно, громко. Звонкостѣ, стп. с. ж. Яркость звучного шѣла. Звонница, цы. с. ж. старин. Колокольня, башня сѣ колоколами. Звонкѣѣ, нѣдѣ с. м. Звонкѣѣѣ, чка, ум. Звонѣцѣ, колокольчикѣ, побрякушка. Звон-</p> <p>3</p>	<p>Навалѣкѣю, кѣши, навалѣкѣѣ, кѣхѣ, влѣбѣ, шѣлѣ. Сл. въ общемѣ же употреб- леннѣ: <i>Навалѣкѣваю</i>, <i>лѣкѣвѣшь</i>, навалѣкѣѣ, навалѣчѣ, навалѣчѣшѣ. гл. л. 1) Множество чего напѣскиваю, изношу. 2) Въ отношенѣ кѣ обла- камѣ употребленнѣся безлично, и зна- чѣннѣ: облака густѣють, и зашѣмѣ- вають свѣтѣ. <i>На нескѣ навалѣкло.</i> <i>Навалѣкою глѣбѣ, злѣбѣ, негодование.</i> Подѣю причину на себѣ гѣвѣшься, негодовать. <i>Трапецѣѣ чашѣ зрѣслу-</i> <i>шнѣнѣѣѣ своимѣѣ не токмо на себѣ,</i> <i>но и на есе лѣпѣтѣсто навалѣкѣѣ</i> <i>Божѣ неслѣговѣленѣе.</i> <i>Навалѣкѣнѣе, нѣдѣ. с. ср. Сл. Исполнен-</i> <i>ное дѣйствѣе навалѣкѣнѣе.</i></p>
	

Таблица 1. Перекосы (искажения) страниц.

3. Проверка контрастности фона. Для этого необходимо использовать инструмент «пипетка» посмотреть цвет фона в произвольном мес-

те. Цвет фона должен быть белым (RGB = 255, 255, 255). Если необходимо увеличить контрастность.

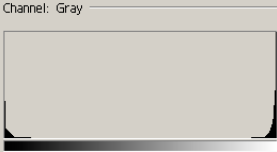
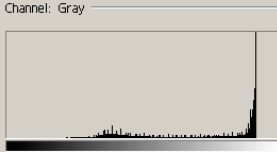
А) высокая контрастность	Б) низкая контрастность
<p>745 ВЛ.</p> <p><i>Сейтокъ, шка. Сейтокекъ, чка. ум. с. и. Сл. Прежде, нежели нынѣшній образъ переплетанія книгъ найдень былъ, листы бумаги приклеивались концами одинъ къ другому, и дѣлали столь длинной столбецъ, какового шребовало пространство сочиненія; столбцы сии завершывали на палку, начиная съ конца сочиненія, а не съ начала. Сверченной такимъ образомъ столбецъ, на которомъ внѣшняя сторона оставалась блѣдалась свитокъ. <i>Принтокъ новъ, великъ, и не мѣлъ ласало мѣлъ.</i> Исаія VIII. 1.</i></p> <p>БЪЛ</p>	<p>ЗВЪЗ.</p> <p><i>Звѣздъ щастливая. Звѣздъ злощастная. Реч. шѣхъ, кои мяхъ, что звѣзды имѣють вліяніе на сложеніе и на судьбу человѣческую, и по гаданіямъ своимъ сказываютъ, кто рожденъ подъ щастливую или злощастную звѣзду.</i></p> <p><i>Звѣздъ во лбу у лошади, есть блѣопяшно.</i></p> <p><i>Звѣздка, дки. с. ж. ум. Подобіе звѣзды живописное, изъ метала вѣданное; вышшо. Табакерка со <i>Ткамъ со звѣздкамъ.</i></i></p> <p><i>Звѣздоска, ки. с. ж. Знаковой и въ семь словъ къ означенію словъ смисль.</i></p> <p>КШО</p>
<p>Channel: Gray</p>  <p>Mean: 237,33 Level: Std Dev: 63,73 Count: Median: 255 Percentile: Pixels: 1589874 Cache Level: 1</p>	<p>Channel: Gray</p>  <p>Mean: 231,72 Level: Std Dev: 32,18 Count: Median: 234 Percentile: Pixels: 1589874 Cache Level: 1</p>

Таблица 2. Контрастность страниц.

4. Увеличение контрастности изображения: Brightness: 0-10; Contrast: 0-20.

5. Обработка изображения, используя фильтры размытия Gaussian Blue (Radius <= 0.3 pixel) так чтобы края букв были размыты.

6. Чистка изображения

Для этого можно использовать фильтр Dust and Scratches «Пыль и царапины» (Radius = 1 pixel).

Чистка изображения, используя инструмент «ластик», удаление грязи, лишних черных точек.





А)	Б)
<p><i>Будябъ</i>, на. с. м. Название порицательное, присвоенное человеку забывчивому, леракому, наглому, мошенному. <i>Какой онъ будябъ.</i></p> <p>Обуяю, или <i>Обудяю</i>, яешь, обуяль, обуять, даяшь, ши. гл. ср. Сл. 1) Обесумяюся, глупымъ дѣлаюся. <i>И обуяюбъ отъ лица меса, его же азъ долю.</i> Перем. ххв. 16. 2) Въ знаменованіи гл. д. Ума лишаю. <i>Не обуи ли Богъ прелюдность мѣра сего?</i> 1. къ Кориня. 1. 20. <i>Мелъ обуябъ страхъ</i> чѣ смыслъ *. Повреждаю, измѣняюся, шерляю <i>соли обуяетъ, гильб</i> в. 13.</p> <p><i>Обуяіе</i>, нѣя. с. ср. измѣненіе.</p> 	<p>ЗИПУНЪ, пунѣ. с. м. <i>Зилучишкѣ</i>, чика. уменьш. простон. Крестьянской кафтанъ, изъ толстаго сержагаго сукна.</p> <p><i>Зилучишко</i>, шка. с. ср. унничк. Худой, Арняной зипунѣ.</p> <p>ЗИ.</p> <p>ЗИЮ, еши, яхъ, зинуши, зѣши. гл. ср. Сл. 1) Отверзаю ротѣ. Говорится о зѣряхъ. <i>Зіаюбъ телюсти. Зілетѣ левѣ.</i> 2) * Готовѣ на поглощеніе кого, на пограніе чего.</p> <p><i>Орлы на тое не азиратѣ</i> <i>То лавозы телюсти</i></p> <p><i>Зіаіе</i>, нѣя. с. ср. Сл. Раздѣстей, паштн. <i>Зіаіе</i> 3Л.</p> <p>ЗЛАКЪ, (злакъ), кя. с. м.</p> 
В)	Г)
<p><i>Сѣдломѣй</i>, мая, моє. <i>Сѣдлобъ</i>, ма, мо. прил. Назвѣщенный, знаемый, вѣстный. <i>Сѣ дѣла сѣу довало сѣдломѣй.</i></p> <p><i>Несѣдломѣй</i>, мая, моє. прил. Тоже что <i>Несѣдломѣй</i>. <i>Сѣе отъ рѣки утасное и Ангелюбъ невадоное танцство, тобою Босородице, на земли ливля.</i> Богор. 4 гласѣ.</p> <p><i>Соевѣсти</i>, сши. с. ж. Врожденная души наша сила или способность судить нравственную доброту или худобу нашихъ дѣяній. <i>Законъ въ сердцахъ своихъ</i></p> <p>2) Внутреннее, внутреннее, нравственное, или худобы дѣла.</p> 	<p>вѣднштой; вѣтви попеременно входящія; листья противоположенные, узкіе, кошенидные. Вѣтви выходятъ изъ пазухъ, исодержашъ по большой часни по три цвѣточка; стбелъки цвѣточные продолговатыя, имѣющіе погнѣзду листочковъ. Цвѣточная чашечка состоишъ изъ пяти узенькихъ листочковъ; цвѣтки овальныя, изъ пяти голубыхъ лепестковъ, едва ноготками соединенныхъ, овальныхъ, остроколючныхъ состоишъ. Въ прочемъ на сши травы сѣя гладкія.</p> <p>чинал отъ Енисея по полямъ Сибиріи по полямъ доновашымъ въ шако что подъ ихолѣ Ав</p> 

Таблица 3. Плохое качество оригинала (грязное изображение).

Формирование исправленной версии изображения страницы.

Интерфейс компоненты гиперграфической информационной системы САР (ФК САР)

Интерфейс окна ФК САР выполнен в виде раскрытой книги. Элементы окна являются сложными графическими образами. Окно содержит следующие элементы: окно правой и левой страниц, кнопки вперед и назад, нижнее меню, титульная надпись.

Окно правой и левой страниц книги предназначено для отображения изображений страниц САР.

Кнопки «вперед» и «назад» предназначены для реализации функции листания страниц. Кнопки выполнены в виде изображений.

Нижнее меню содержит следующие пункты: «главная», «поиск», «выход», позволяет перейти в главное окно, окно поиска по страницам словаря или выйти из программы.

Титульная надпись информирует пользователя о том, какой из томов словаря загружен.

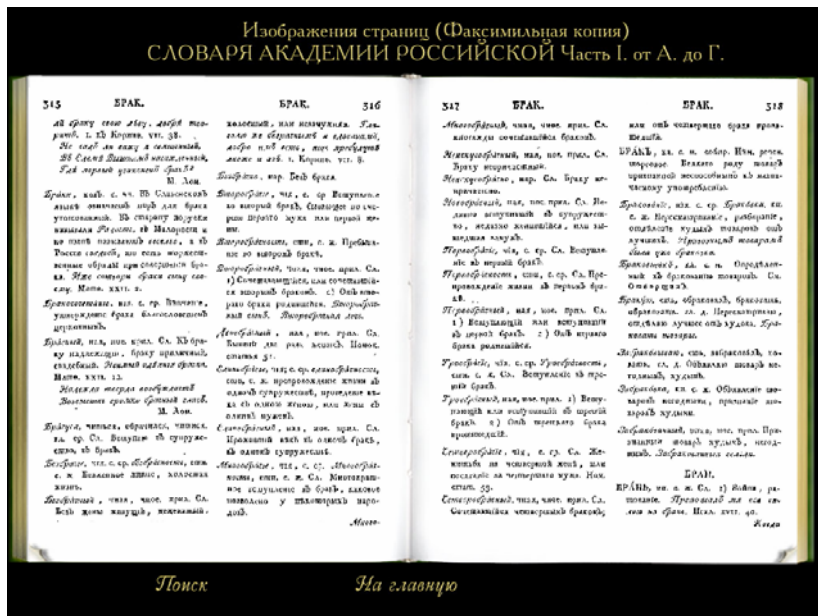


Рис. 2. Окно ФК САР.

Работа в окне ФК САР

Окно ФК САР предназначено для работы со старой книгой, а точ-

нее с изображениями ее страниц. Пользователь имеет возможность листать страницы книги. Для этого необходимо кликнуть мышью на изображения загнутой страницы книги в нижнем правом нижнем углу страницы для перемещения вперед или в левом нижнем углу левой страницы для перемещения назад. Если пользователь находится на первой или последней странице, то данная функция будет недоступна соответственно для перемещения назад или вперед.

Для осуществления поиска по страницам необходимо кликнуть по соответствующему пункту в нижнем меню. При этом откроется окно поиска по страницам.

Окно содержит информационную надпись, строку для ввода условий поиска и кнопку «Найти», по нажатию которой осуществляется переход в окно факсимильной копии САР и поиск соответствующей страницы.

Для перехода в главное окно необходимо кликнуть мышью по соответствующему пункту в нижнем меню.