ЦИФРОВАЯ БИБЛИОТЕКА GREENSTONE

ОТ БУМАГИ К КОЛЛЕКЦИИ

Dr Michel Loots, Dan Camarzan and Ian H. Witten

Human Info NGO, Belgium
Simple Words, Romania
University of Waikato, New Zealand

Органайзер коллекции, проще Органайзер, является свободно доступным пакетом программ, предназначенным для помощи пользователю создавать и редактировать материал, связанный с коллекцией. Он распространялся с более ранними версиями Greenstone. Его функциональные возможности были заменены Библиотечным интерфейсом, описанным в *Руководстве пользователя цифровой библиотеки Greenstone.* Этот документ, который обеспечивает обратную совместимость, описывает, как использовать Органайзер.

Мы надеемся, что это программное обеспечение работает хорошо. Пожалуйста, сообщите о любых проблемах по адресу: *greenstone@cs.waikato.ac.nz*

Greenstone gsdl-2.50 **Март 2004**

Об этой инструкции

Эта инструкция детально объясняет, как создавать CD-ROM коллекции из бумажных документов. Здесь детально описываются процедуры и экономические стороны процесса сканирования и оптического распознавания символов (OPC), так что вы переведете текст в правильный формат, применяя программное обеспечение Greenstone. Это также описывает, как создать и редактировать материал, связанный с собранием.

В нашем объяснении мы старались быть ясными насколько возможно. Ссылка на любые торговые марки или продукты компании – использовались для иллюстративных целей, и не подразумевают, что в сравнении с любым другим , мы рекомендуем или одобряем это изделие.

Сопутствующие документы

Полный комплект документации к Greenstone состоит из пяти томов:

- Руководство по установке цифровой библиотеки Greenstone
- Инструкция для пользователя цифровой библиотеки Greenstone
- Руководство разработчик а цифровой библиотеки Greenstone
- Цифровая библиотека Greenstone: от Бумаги до Коллекции (этот документ)
- Цифровая библиотека Greenstone: Использование Органайзера

Благодарность

Операции сканирования и другие ноу-хау связанные с созданием совместных не коммерческих коллекций были разработаны Майклом Лотсом, МД НПО Хьюман Инфо и ХьюманСД, Дэном Камарзаном из Симл Волд и его группой в сотрудничестве с Брасов, Румыния.

Программное обеспечение Greenstone - продукт совместного труда множества людей. Rodger McNab и Stefan Boddie принципиальные разработчики системы. Неоценимый вклад внесли David Bainbridge, George Buchanan, Hong Chen, Michael Dewsnip, Katherine Don, Elke Duncker, Carl Gutwin, Geoff Holmes, Dana McKay, John McPherson, Craig Nevill-Manning, Dynal Patel, Gordon Paynter, Bernhard Pfahringer, Todd Reed, Bill Rogers, John Thompson, и Stuart Yeates. Остальные члены Проекта Новозеландской цифровой библиотеки разработали вдохновенный дизайн всей системы: Mark Apperley, Sally Jo Cunningham, Matt Jones, Steve Jones, Te Taka Keegan, Michel Loots, Malika Mahoui, Gary Marsden, Dave Nichols и Lloyd Smith. Мы также выражаем свою признательность всем тем, кто трудился над созданием пакетов, попадающих под действие лицензии GNU, и включенных в дистрибутив: MG, GDBM, PDFTOHTML, PERL, WGET, WVWARE и XLHTML.

Содержание

1 Введение

Одна из целей создания программного обеспечения Greenstone состоит в том, чтобы представить возможность различным институтам, организациям, агентствам ООН, неправительственным и некоммерческим организациям, а также правительствам создавать информационные коллекции, которые можно расположить как в Интернете, так и сохранить на CD-ROM.

i. Ниже описана обычная процедура при создании коллекций:
ii. Подбор исходных документов
iii. Получение разрешения владельцев авторских прав на использование этих документов.
iv. Сканирование и оптическое распознавание символов (ОРС) материальных документов для их преобразования в цифровой формат
v. Преобразование всех документов в единый формат, который можно импортировать в Greenstone (желательно HTML и Microsoft Word, но можно также и в другие форматы) с помощью дополнений к программе "plugin" (смотрите Инструкцию Пользователя).
vi. Маркировка глав, параграфов и рисунков в цифровых документах.
vii. Организация коллекций в оптимально структурированную цифровую библиотеку.
viii. Построение цифровой библиотеки.
ix. Запись и распространение коллекций на CD-ROM и/или их публикация в Интернете.

Для того, чтобы создать цифровую коллекцию, все публикации соответственно нужно преобразовать в цифровой формат. Если доступны только твердые копии книг, документов, то их необходимо отсканировать и перевести в форму, считываемую компьютером (ш шаг). Обычно это делается путем проведения оптического распознавания символов, в некоторых случаях - простым перепечатыванием. Этот процесс рассматривается в 2-4 главах.

v. шаг позволяет выбирать отдельные части документа и располагать их в определенном порядке в библиотеке, vi шаг включает в себя присвоение определенных атрибутов документам, таких как указатели названий, ключевые слова и библиографические данные для осуществления запроса и поиска по библиотеке. Эти шаги

рассматриваются в пятой главе.

Эта инструкция также обсуждает множество вопросов, касающихся процесса редактирования при создании цифровых коллекций из твердой копии. Перед тем, как продолжить чтение, мы рекомендуем вам ответить на следующие вопросы:

- Какова цель создания вашей коллекции?
- На какую группу людей она ориентирована?
- Насколько она масштабна — локальна, региональна или глобальна?
- Сколько содержит документов?
- Сколько содержит страниц?
- Сколько содержит графических объектов?
- Имеет ли коллекция части, предназначенные для ограниченного круга людей, а другие - для всех пользователей?
- Доступны ли исходные документы в электронном формате?
- Если да, то в каком? (нужно отметить, что PDF-файлы не являются эквивалентами полнотекстового цифрового формата, так как являются лишь копиями страниц)
- Каков статус авторских прав исходных документов?
- Кто владеет авторскими правами?
- Существуют ли другие организации, располагающие продуктом, ориентированным на ту же самую группу людей?
- Желаете ли вы работать вместе с другими группами?
- Каков бюджет, выделенный на этот проект?
- Какие доступны человеческие ресурсы(в человеко-месяцах) для координации, редактирования, сканирования и программирования?
- Сколько доступно компьютеров?
- Сколько CD-ROM вы хотите создать?
- Будут ли они бесплатными или платными?

2
Сканеры и сканирование

Первый шаг для преобразования бумажных документов в цифровую коллекцию состоит в сканировании всех страниц исходных документов. Следующий шаг - это проведения процесса оптического распознавания символов (OPC), для чего очень важны высококачественные и четкие исходные документы. Процесс перевода в цифровой формат нуждается в сканере, способном работать при разрешении 300 dpi (точек на дюйм). Большую часть сканирования можно произвести в черно-белом режиме, но при включении цветных иллюстраций их следует отсканировать цветным сканером. В большинстве случаев обложки книг являются красочными и их необходимо сканировать в режиме цветного рисунка.

2.1 Сканеры

Сканеры доступны по различным ценам и имеются всевозможных размеров и форм. Их цена находится в пределах от $ 100 за обычный настольный сканер до $50000 за огромные индустриальные сканеры таких производителей как Bell & Howell[1] . Существует много разных торговых Интернет-страниц, предлагающих всевозможные сканеры. Для поиска сканеров просто используйте такие поисковые системы, как Google, Altavista, Yahoo.

Обычный формат, в котором сохраняется отсканированный документ, это TIFF или BMP (Bitmap image). Сжатая форма TIFF IV является лучшим форматом для использования. Средняя отсканированная страница, переведенная в этот формат, занимает всего 50 килобайт, по сравнению с 2 Mb идентичной страницы, сохраненной в BMP.

Недорогие настольные сканеры (flat-based scanners)

Настольные сканеры являются самыми дешевыми и наиболее доступными сканерами. Существует много торговых марок, таких как HP, Agfa, Acer и т.д. Их цены колеблются от 100 до 300 долларов. С их помощью можно сканировать как черно-белые рисунки, так и цветные. Низкие цены позволяют каждому

[1] Все цены, упомянутые в этом документе, даны в долларах США, 2001 года.

пользователю иметь один из таких сканнеров.

Их недостатки - это средний уровень качества, медленное сканирование, ненадежность при высоких температурах и относительно частые поломки. Страницы должны сканироваться вручную и каждая из страниц по отдельности. Каждую страницу нужно положить так, чтобы она располагалась правильно. Продуктивность на таких сканерах очень низка. Несмотря на то, что производители утверждают, что одну страницу можно отсканировать меньше, чем за одну минуту, на практике трудно преодолеть границу в 12 страниц за час. К тому же процесс сканирования полностью занимает компьютерные ресурсы.

Следовательно, такие сканеры удобны только для выполнения небольших работ с небольшим набором страниц - не более 200-400 в месяц, если выполнять сканирование регулярно. Те, кто работает со сканером полный день, это составляет 1000-2000 страниц.

Сканеры с автоматической подачей бумаги

Сканеры с автоматической подачей бумаги стоят около 500-1200 долларов. Можно помещать до 10-15 страниц в сканер и сканировать их одновременно: следовательно, оператору не нужно постоянно подходить к аппарату. Это позволяет увеличить производительность до 150-200 страниц в день. Эти сканеры наиболее стойкие и не нуждаются в ремонте в течение долгого времени (после сканирования 30000-50000 страниц).

Их недостаток состоит в том, что сканируется только одна сторона страницы; для того, чтобы отсканировать другую сторону, страницу нужно перевернуть. Это часто создает проблемы, так как автоматическая подача бумаг всегда проблематична и часто страницы застревают.

Эти сканеры удобны для обработки 1500-3000 страниц в месяц.

Цветные сканеры

Любая операция сканирования сталкивается с цветными рисунками, поэтому необходимо наготове иметь цветной сканер. Обычно менее 5% любой публикации содержит какие-либо цветные рисунки, включая обложку. Поэтому рекомендуется иметь дешевый цветной настольный сканер. Рекомендуется иметь сканер с разрешением не менее 600 dpi.

Профессиональные дуплексные сканеры

Профессиональные сканеры надежны, прочны и способны сканировать около 2000-10000 страниц в день. Они имеют автоматическую систему подачи бумаги, которая обрабатывает стопки в 50-200 страниц. Самые лучшие и быстрые сканеры - это дуплексные сканеры, которые сканируют обе стороны документа одновременно.

Эти сканеры нуждаются в мощном компьютере с жестким диском как минимум на 10-20 Gb. Их цены граничат от $5000 до $50000. Например, дуплексный сканнер Canon DR-6020 стоит 5000 долларов и может работать с двухсторонними документами. Он способен сканировать до 2000 в день и находится в рабочем состоянии до 600000-800000 страниц. Сканеры Bell&Howell and Fujitsu стоят от 10000-50000 долларов и могут обработать миллионы документов до первого ремонта.

Сканеры Micro-fische стоят от $15000 за полуавтоматический набор до $80000, работающий полностью в автоматическом режиме.

Программы для сканирования

Каждый сканер имеет свое программное обеспечение, поэтому эту программу необходимо установить на ваш компьютер. Некоторые программы имеют computer card, которую нужно установить для ускорения процесса сканирования.

2.2 Подготовка документов

Документы нужно подготовить заранее до их сканирования. С них нужно стряхнуть пыль, высушить мокрые документы, снять скрепки и расправить страницы.

Необходимо аккуратно снять сшивку с каждой книги. Многие книги, которые вы будете использовать для сканирования, необходимо будет снова сшить, поэтому будьте внимательны при снятии сшивки. Для документов, имеющих объем более 20 страниц, мы рекомендуем поручить задание на сканирование специализированным организациям, имеющим соответствующее оборудование.

2.3 Процесс сканирования

Используя программу, распространяемую вместе со сканером, после сканирования вы получите цифровой документ в формате Bitmap

или TIFF. Эти документы нужно сохранить на жестком диске под любым именем.

Контроль качества

Конечная цель сканирования - это либо провести отсканированные документы через процесс оптического распознавания символов OPC (optical character recognition) для получения документа в формате Word или HTML, либо получить изображения документов в формате PDF. В любом случае качество изображения исходного документа очень важно. Если качество изображения ниже стандарта, то они будут выглядеть размыто и занимать больше компьютерной памяти. Качество изображения также очень сильно влияет на процесс оптического распознавания символов (далее OPC). При качестве ниже нормы его продуктивность падает на 40%. OPC, как правило, составляет 90% от общей стоимости всего процесса преобразования твердых копий в цифровой формат, поэтому качество сканирования может очень сильно повлиять на конечные затраты.

Качество файлов формата TIFF можно улучшить путем настройки процесса сканирования для каждого типа бумаги, изменяя настройки в программе. Относительно тонкая бумага потребует других настроек, нежели другая: контраст должен быть настроен в зависимости от качества печати документа, который будет отсканирован , и т.д.

Вначале разделите исходный материал на группы со схожим качеством бумаги и схожим качеством печати. Проведите испытание OPC на нескольких экземплярах бумаг из первой группы для определения оптимальных настроек. Затем можете смело сканировать весь материал из этой группы.

Рекомендуемые правила для обозначения документов

Присвойте каждой книге или документу свой номер или код, который станет названием папки, содержащей все изображения TIFF из этого документа. В зависимости от операционной системы (DOS, Windows, UNIX, LINUX и т.д.) можно использовать от 8 до 128 символов в названии файла. Мы рекомендуем не превышать 8-16 символов. Первые пять букв могут символизировать название документа, следующие буквы языковой код, а остальные цифры - определенную страницу). Например, название *u7548el2.tif* может означать рисунок TIFF на странице 12, написанной на английском языке под кодом *u7548e*.

Создайте одну директорию на жестком диске для отсканированных работ, назовите, к примеру, *scanjobs*. Затем создайте поддиректорию для каждого задания. Внутри поддиректории создайте еще одну поддиректорию для каждой публикации и книги - и7548 в примере, показанном выше. Сохраняйте все изображения TIFF из документа, включая цветные рисунки, в этой папке.

2.4 Производительность и ресурсы

Вы не должны недооценивать весь масштаб процедуры сканирования — в особенности процесс ОРС. Лучше всего считать сканирование и ОРС как полностью независимые операции. Оптимальное решение должно быть принято по каждому из них в отдельности.

Вот некоторые вещи, о которых следует подумать перед приобретением сканеров и компьютеров: наличие необходимого помещения и рабочей силы, обучение рабочих; зарплата; минимальное и максимальное число страниц, которые необходимо отсканировать; сроки; можно ли эти документы передавать третьим лицам.

Стоимость сканирования

Важное решение, которое нужно принять, - покупать ли сканирующую аппаратуру и проводить сканирование самим, или поручить это дело компании, специализирующейся на сканировании. Вот главные пункты, над которыми надо подумать:

- Конечные сроки для завершения сканирования
- Общее количество страниц
- Зарплата тех, кто совершает сканирование

Люди, занимающиеся сканированием, должны иметь высокую мотивацию, быть искусными и ответственными за качество работы.

Средняя цена за одну отсканированную страницу у профессиональных компаний составляет $0.06. К этому нужно добавить стоимость доставки, которая может достигать $0.03 за транспортировку страницы из развивающейся страны в развитую, и $0.015 - за страницу в самой стране.

В таблице 1 приведена стоимость выполнения операции сканирования вашими усилиями с использованием разных типов сканеров. Заметьте, что все цифры приблизительны. Они примерны

и основываются на опыте авторов. Первые три колонки включают стоимость рабочей силы. Первая из них - это производительность в страницах/месяц при работе на полную ставку. Расчет человеко-часов на страницу производится путем разделения числа рабочих часов в месяце на производительность страница/месяц и представлен во второй колонке. При расчете предполагается 180 рабочих часов в месяц.

Таблица 1 Стоимость сканирования

	Объем (страниц/месяц)	Час/страница (180 часов в месяц)	Цена/стр (предполатая $4/час)	Цена сканера	Работоспособность до первого ремонта (страниц)	Объем выполняемый профессиональной компанией (при $0,06 за страницу)
Страничный настольный сканер Сканер с автоматической подачей страниц	2,500	0.072	$0.288	$300	7,000	5,000
Профессиональный low end duplex	8,000	0.0225	$0.09	$800	30,000	13,000
Профессиональный: high-end duplex	40,000	0.0045	$0.018	$6,000	600,000	100,000
	150,000	0.0012	$0.0048	$50,000	8,000,000	833,000

Чтобы определить стоимость сканирования одной страницы, умножьте полную часовую зарплату на вторую колонку. К примеру, в третьей колонке представлена цена выполнения сканирования при найме труда без учета затрат на сканер - $4/час.

Эти подсчеты подразумевают, что сканер будет использован для достаточно большого количества копий, чтобы окупить свою начальную стоимость. Последние три колонки дают больше информации о стоимости самого сканера. Первая из них показывает стоимость сканера, следующая приблизительную продолжительность работоспособности. Последняя показывает число страниц, которые можно отсканировать, наняв компанию, при

цене $0.06/страница.

Конечно же на выбор сканера влияют и другие факторы: наличие денег, необходимость в минимизации зависимости от других, договоренность с библиотекой, разрешающая осуществлять сканирование книг, не вывозя их за пределы библиотеки и т.д.

Таблица приведенная выше, дает примерную оценку количества страниц, которое необходимо отсканировать для окупаемости затрат. Очень редки случаи, когда организация нуждается в сканировании 800000 страниц. При таком масштабе появляются более сложные проблемы, такие как содержание оборудования и окупаемость затрат путем сдачи его в аренду, которые не будут обсуждаться в этой части.

Развитие бизнеса сканирования может показаться очень привлекательной коммерческой возможностью, в особенности в развивающихся странах. Но помните, Что после того, как будут отсканированы документы, ваши клиенты больше никогда не закажут сканирование тех же самых документов - вне зависимости от того, насколько хороши ваши с ними отношения. С коммерческой точки зрения этот бизнес нуждается в интенсивном маркетинговом изучении. Мы не советуем неправительственным организациям и некоммерческим организациям заниматься таким бизнесом без детального исследования рынка и хорошо продуманного бизнес-плана.

В заключение отметим, что если нужно сканировать от 10000 до 50000 страниц, лучше поручить эту работу сканирующей компании. Профессиональный low-end сканер будет окуплен только в том случае, если вы отсканируете более 100000 страниц. Если вы решили приобрести такой сканер, то лучше это сделать совместно с другой неправительственной организацией или библиотекой.

3
OPC: оптический распознаватель символов

Оптический распознаватель символов или OPC трансформирует отсканированное цифровое изображение в текст. Исходный материал - это цифровое изображение в формате TIFF или Bitmap — желательно чтобы он был хорошего качества. После прохождения через OPC вы можете получить файл в формате RTF, Word, HTML на ваше усмотрение.

Вот шаги, используемые при переводе бумажных документов в цифровой формат:

1. сканирование
2. анализ
3. распознавание
4. сканирование изображений и таблиц.

Следуя им, вы можете проводить контроль качества полученных файлов и сохранять их в соответствующей папке.

На рынке существует достаточно много хороших программ OPC стоимостью от $100-400. Вот несколько из них:

- Read-Iris (http://www.readiris.com/)
- Omnipage (http://www.omnipage.com/)
- Fine-Reader (http://www.finereader.com/)

Вся информация, включая перечень дистрибьютеров, находится на Интернет сайте производителей. Среди них, по опыту автора, самые легкие в использовании Fine-Reader и Omnipage. Fine-Reader является самым дешевым и стоит всего $100. Он предлагает гибкие возможности и наибольшее количество разных языков.

Вам нужно решить, проводить ли сканирование и OPC своими усилиями или поручить это компании, специализирующейся в этой сфере. Для того, чтобы провести этот процесс своими усилиями, вам нужен сканер, программное обеспечение OPC, развитие навыков в работе с OPC, работники, нацеленные на качество исполнения работы.

3.1 Процесс OPC

Процесс OPC различается во всех программах OPC, и любая из выбранных программ требует добротного изучения. Инструкция по эксплуатации каждой программы объясняет этот процесс в деталях. Четыре пункта процесса OPC заслуживают особого внимания: контроль качества, таблицы, изображения и специфический материал - такой, как формулы, иностранный язык и т.д.

Контроль качества

Мы снова и снова хотим заострить ваше внимание на контроле качества. Контроль качества лучше поручать людям, чей родной язык является таким, на котором написан документ, или они владеют этим языком в совершенстве. Лучшие работники - это школьники и студенты, потому что молодые люди более внимательны и более сконцентрированы при таком виде работы, нежели пожилые люди.

Обычно существует четыре этапа контроля качества.

Первый проводится одновременно с процессом OPC. Каждая программа OPC имеет встроенную проверку орфографии, которая подчеркивает подозрительные по написанию слова.

Второй этап - общая проверка текста после завершения процесса OPC. Очень часто встречаются такие ошибки, как пропуск страниц, абзацев, названий глав и т.д. Нужно провести общий обзор на наличие всех страниц и проверку заголовков, названий глав, абзацев и таблиц.

Третий этап - проверка орфографии в Microsoft Word. В Word, как правило, орфографические словари более исчерпывающие, чем в программах OPC. Путем импорта документа в Word и проведения орфографической проверки можно определить и исправить большее число ошибок. Не забудьте ввести в параметры проверки комплексные слова и термины, которые присущи такого рода текстам, на наличие в них ошибок.

И наконец, на четвертом этапе завершенный документ должен быть проверен другим человеком, занимающимся составлением полной книги, проверяющим орфографию, наличие проблем с таблицами, изображениями и общим видом завершенного текста. Только после этого электронную книгу можно распространять.

Таблицы

Программы OPC плохо справляются с таблицами. Более того, таблицы трудно проверять. Они содержат много цифр и такие символы, как точки, запятые, которые по ошибке могут быть перенесены программой в другой столбец или строку. Поэтому при проверке необходимо внимание, упорная работа, терпение и контроль качества. С ними можно работать тремя обычными способами.

Первый состоит в том, что можно работать с таблицами как с изображением. Это включает в себя черно-белое сканирование изображений таблиц и перенос их в этой форме в нужное место документа. Это самый легкий путь. Не будет никаких ошибок, и все потраченное время уйдет только на создание изображения. Однако, полученные изображения таблиц будут занимать больше компьютерной памяти. Также разрешение экрана компьютера может быть недостаточным при выводе на экран больших таблиц. Если вы захотите разместить всю таблицу на экране, то разрешение экрана может быть недостаточным. Если таблица слишком широка, то пользователю придется просматривать все колонки и строки, не видя их названий.

Второй метод состоит в создании таблиц вручную, определив нужное количество строк и столбцов, и вручную впечатать в них данные.

Третий способ заключается в том, что таблицу можно провести через OPC. Это сохранит больше времени, чем второй метод, но потенциальное количество ошибок увеличится. Некоторые колонки иногда сливаются, а иногда программа не распознает точки и запятые.

Изображения

Документы содержат три основных вида изображений:

- Черно- белые рисунки
- Черно-белые фотографии
- Цветные фотографии

Черно-белые рисунки следует сканировать в штриховом режиме (line art mode) и желательно сохранять в виде файлов GIF или PNG. Чернобелые фотографии следует сканировать в режиме grayscale и сохранять как файлы GIF или JPEG. Цветные фотографии нужно

сканировать в цветном режиме и сохранять как файлы JPEG. Обычно файл JPEG среднего качества имеет необходимое разрешение.

Многие рисунки занимают наибольшее пространство в коллекции, сохраненной на жестком диске или на CD-ROM. Поэтому очень важно оптимизировать настройки изображения, сделав его как можно разборчивее и четче, в то же время уменьшая его размер. Для сохранения места предпочтительнее не включать некоторые изображения, не относящиеся к тексту.

Рисунки должны быть отсканированы каждый по отдельности. Мы рекомендуем называть эти рисунки первыми 6 буквами названия документа, а остальными цифрами номер страницы, на которой он располагается. В качестве альтернативы, предполагая, что каждый документ находится в своей директории, можно просто использовать букву, после чего идет страница, на которой находится это изображение. Если на одной странице существуют несколько рисунков, то соответственно используйте другие буквы для их обозначения. Например, если изображение в формате JPEG появится на странице 36 публикации *u7548e*, то оно будет помещено в файл названный *u7548e36.jpg* или *p36.jpg*.

После того, как изображения отсканированы, вы можете использовать специализированные программы для редактирования размера и расположения рисунка.

Специализированный материал

Многие документы содержат технические термины, такие как специализированные символы, формулы и неразборчивые страницы. Эти трудно распознаваемые символы, как правило, связаны с разными язьпсами. Для каждого документа вам необходимо выбрать соответствующий язык в опциях программы OPC. Формулы нужно будет перепечатывать вручную, так как во многих случаях OPC их не распознает, и их необходимо вводить в Word. Неразборчивые страницы могут содержать материал, который трудно воспроизвести из-за каких-либо повреждений и т.д., поэтому такие места придется перепечатывать.

3.2 Производительность и доступные ресурсы

Как было упомянуто ранее, вы не должны недооценивать сложность процесса OPC. Хотя процесс OPC нужно рассматривать отдельно от сканирования, для его оценки применяются схожие практические

рекомендации: необходимые ресурсы для приобретения компьютеров, доступность рабочей силы и умение руководить; зарплата; общее количество страниц, нуждающихся в обработке; можно ли передавать эти документы третьим лицам.

В следующим разделе мы поделимся с вами нашим опытом работы с OPC в таких странах, как Бельгия, Индия и Румыния. Все исследования, подсчеты и результаты выполнены для типичных условий - документы средней сложности (включая таблицы и изображения), которые встречаются в большинстве библиотек и архивов, высокое качество результатов и средняя-долгая длительность работы.

Интенсивный OPC

Процесс OPC является трудным. Он требует большой концентрации внимания и умения. Перед попыткой достигнуть максимальных результатов требуется порядка 6 недель работы, в процессе которой идет нарастание опыта.

Обычно максимальная производительность достигается в первые часы начала дня. После трех часов работы с OPC она резко падает, примерно на 50% от начального уровня. После 6 часов большинство людей очень устают.

То же самое происходит в течение первых недель работы. В первые недели продуктивность находится на высоком уровне, но после этого 2/3 людей устают и теряют интерес к работе. Такие люди либо уходят, либо продолжают работать на очень низком уровне, влияющем на качество и производительность. Даже те, кто выстоял критические 1-5 недель и становится частью рабочей команды, часто уходят в поисках лучшей работы в промежутке 6-12 месяцев.

Заметки, которые были сделаны в части 3.1 о рабочей силе, относятся в особенности к интенсивному OPC. Контроль качества лучше поручать людям, чей родной язык является тем же, на котором написан документ или они владеют этим языком в совершенстве. Лучшие работники - это школьники и студенты, потому что молодые люди могут быть более внимательны и более сконцентрированы при таком виде работы, нежели более взрослое поколение и пожилые люди. Условный критерий отбора таков - люди в возрасте от 18 до 23 лет обычно лучше подходят на такую работу, нежели те, кому больше 25.

И наконец, процесс OPC очень рутинен и скучен. Поэтому нужно

как-то поддерживать рабочий дух, мотивацию и привязанность к работе.

Вот итог вышеперечисленного:

- Молодые люди в возраста 18-25 лет самые лучшие кандидаты на такую работу.
- Так как первые часы работы являются самыми продуктивными, следует нанимать рабочих на пол-дня, и только самым усердным работникам позволять работать целый рабочий день.
- Две трети работников покидают работу от усталости или от рутины в течение 3-5 недель, все это отражается на ухудшении продуктивности в этот период.
- Нужен постоянный приток работников для покрытия затрат на обучение, поддержание усердия и рабочего духа.

Таблица 2 продуктивность процесса ОРС

	Рабочие часы/день	Страницы/день	Страницы/месяц
Начальная подготовка *(6* недель)	3	6	120
Оптимальный уровень продуктивности	3	9	150 to 200
	7	28	500 to 600

В Таблице 2 приведены средние цифры продуктивности работы на ОРС. Документы приходят разными и по размеру, и по содержанию. При составлении этой таблицы были учтены такие факторы, как разнообразие документов содержащих среднее количество изображений и таблиц - к примеру один рисунок и одна таблица 5 на 5 на каждые восемь страниц. Также предположено, что изображения являются среднего-высокого качества. Заметьте, как это было уже рассмотрено, что это зависит от качества сканирования и также от того, насколько хорошо работники знают язык, на котором написаны документы.

В таблице также приведены данные о тех, кто находится в процессе обучения, и тех, которые уже работают в оптимальном режиме. Если член руководящей команды уделит до трех часов в день на процесс ОРС, то он может достигнуть результата 180-200 страниц в месяц. Для постоянного работника, прошедшего хорошую подготовку, с высокой внимательностью и отдачей достигнуть 500-600 страниц в

месяц не будет проблемой.

Тем не менее, с неразборчивыми документами плохого качества и изобилием таблиц и рисунков эти цифры будут ниже — наверное 300-400 страниц для постоянного работника.

Представьте, что зарплата для мотивированного постоянного работника составляет $400 в месяц, а затраты на менеджмент, компьютеры, аренду, коммунальные услуги и т.д. стоит $300-400 на человека в месяц. Плата за 1 страницу ОРС $ 1.2-1.6. Учитывая подготовительный период, общий объем, время, затраты на увольнение при закрытии бизнеса, эти цифры поднимутся до $1.5-2.5 за страницу.

Стоимостьвыполнения работы самим нужно сравнивать со стоимостью, если поручить эту работу профессиональным ОРС компаниям. Неправительственная организация в Румынии предлагает такие услуги, и цена для гуманитарных некоммерческих организаций ниже и граничит от $1.2 до $2 за страницу. Если у вас есть какие-либо вопросы, пишите нам на scanning@humaninfo.org.

3.3 Альтернатива ОРС

Существуют две альтернативы проведению ОРС.

Ручное перепечатывание

Используя этот метод, можно не использовать сканер вообще, применив простой текстовой редактор. Единственное, что остаётся сделать, это отсканировать обложку и изображения, поэтому можно обойтись без дорогих сканеров и программ ОРС.

Люди, вьшолняющие эту работу, не обязательно должны вникать в суть текста. Все, что от них требуется, это аккуратно и безошибочно печатать то, что они видят. Тем не менее, перепечатывание вносит ошибки, поэтому используется метод двойного набора. Два человека перепечатывают один и тот же текст, после чего специальная программа проверяет обе электронные версии этого документа, слово в слово сравнивая его с оригиналом. Предполагается, что слово, напечатанное одинаково в обоих случаях, является правильным, хотя это не всегда так, поэтому также практикуется и метод тройного набора.

Положительной особенностью использования метода перепечатывания является то, что можно снизить затраты, так как

нет необходимости приобретать программы OPC компьютеры могут быть более старой модификации или б/у, в то время как для проведения OPC нужны мощные компьютеры. К тому же работа может выполняться менее квалифицированными работниками. Один недостаток состоит в том, что необходим подготовительный период, как минимум до двух месяцев. Набор одним человеком приводит к множеству ошибок, поэтому приходится проводить двойной или тройной набор текста, что связано с дополнительными затратами.

Все издержки зависят от уровня зарплаты. Обычно люди, занимающиеся печатанием, в развивающихся странах получают около $150 в месяц. Их производительность может составлять 20-30 страниц в день, соответственно 400 страниц в месяц, включая изображения. С двойным набором это становится $300 в месяц плюс другие затраты.

Файлы Изображения

Очень дешевой альтернативой OPC является использование простого формата PDF для всех отсканированных документов. Цена составляет всего лишь часть от стоимости OPC — около $0.1 за страницу.

После того, как завершено сканирование и доступным файлы TIFF, автоматический конвертер может (обычно Adobe Acrobat, Adobe Photoshop) преобразовать все файлы формата TIFF в файлы PDF.

Отрицательной чертой таких документов является то, что по ним нельзя осуществлять поиск. К тому же они довольно больших размеров, обычно 50кб на страницу плюс 20% в зависимости от качества файла TIFF.

Файлы PDF очень долго загружаются с Интернета (в развивающихся странах это очень дорого и многим не по карману). Они редко помещаются на флоппи-диске и не поддерживают такие необходимые функции, как "вырезать" и "вставить".

Прибегать к использованию PDF-файлов необходимо только тогда, когда нет денег на OPC, и для документов, которые будут использованы относительно малым количеством людей, имеющих быструю Интернет-связь.

3.4 Совмещение сканирования с OPC

Если сканер напрямую подключен к компьютеру, на котором

установлена программа OPC, то большинство этих программ может проводить процесс сканирования и OPC одновременно. Этот метод является хорошей стратегией, если вы работаете с небольшим объемом информации, но займет очень много времени, если он велик.

Если вы желаете придерживаться темпа 100-150 страниц в месяц, то этот метод для вас. Для большего объема документов быстрее и удобнее совершать сканирование отдельно от OPC.

4
Три примера: от 1000 до 100.000 страниц

4.1 Небольшая коллекция: от 500 до 1000 страниц

Большинство НПО нуждаются в сканировании материала объемом 5000-1000 страниц. Это можно осуществить и провести с помощью ОРС без особого труда при наличии высоко целеустремленных работников.

Сканирование

Первый шаг состоит в сканировании материала для получения его в качественном электронном формате TIFF - на каждую страницу по одному файлу. В настройках надо подбирать соответствующие режимы для каждого типа рисунка. Предположим, что задача состоит в том, чтобы отсканировать 1000 страниц. Это можно организовать, наняв работников на полставки на протяжении месяца — только для сканирования. Файлы TIFF займут от 60 до 80MB пространства на жестком диске и лучше всего также записать эти файлы на компакт-диск (CD-ROM). Недорогой настольный сканер ценой $ 100-300 будет достаточен для этой работы. Сканирование можно проводить вечером или по выходным.

ОРС

Следующий шаг - это провести процесс ОРС одним или несколькими работниками, имеющими хорошие языковые навыки. Файлы TIFF можно распределить по нескольким компьютерам или выполнять всю работу на одном. Обычно это займет около пяти или шести месяцев для работника на пол -ставки (примерно 20 часов в неделю) для полной и безошибочной обработки 1000 страниц в файлы Word или HTML-документов.

Поручение другим

В качестве альтернативы вы можете поручить эту работу профессиональной компании. Это будет стоить около $1500-$2000 для получения отличных Word или HTML-файлов.

4.2 Все публикации организации: 5000 страниц

Многие большие организации имеют архивы, состоящие примерно из 5000 страниц изданий, книг, различного рода литературы, которые уже не издаются.

Сканирование

Это слишком большой объем для простого настолько сканера. Сканирование лучше либо поручить компании (около $400 за 5000 страниц) или приобрести сканер с автоматической подачей страниц (около $900). Можно также приобрести сканер совместно с другими НПО или другими организациями ($6000 разделить на количество совладельцев). Все 5000 страниц займут около 300-400 MB, и снова мы рекомендуем вам параллельно записать это на CD.

OPC

И снова вам нужно пройти через процесс OPC, используя схожую рабочую силу, что и в предыдущем случае. Можно использовать один или несколько компьютеров. На 5000 страниц это займет около 25-30 месяцев при работе по 20 часов в неделю. На практике это очень долгий срок. Нужен человек, ответственный за плату работникам, надзором за качеством, предоставление необходимого места и т.д. для окончания работы качественно и за менее короткий срок.

Можно также создать файлы PDF, которые займут около 300-400 MB пространства и будут очень долго загружаться через Интернет.

Поручение другим

Можно поручить другой компании как сканирование, так и OPC. Это будет стоить от $7500 до $10000 за всю работу.

4.3 Небольшая библиотека: 100.000 страниц

Различные крупные организации, университеты, правительство и специализированные библиотеки могут пожелать перевести всю библиотеку в цифровой формат — допустим 100.000 страниц. Первая вещь, о которой следует позаботиться, - это авторские права на все издания. Вы должны получить разрешение владельцев этих прав перед тем, как начать сканирование. Вам также следует убедиться в том, что эти документы еще не существуют в цифровом формате.

Сканирование

Такой объем слишком велик для сканера с автоматической подачей страниц. Это нужно либо поручить другим организациям ($8000 за 100.000 страниц) или приобрести дорогой сканер совместно с другими ($6000). Объем 100.000 страниц займет 6-8 GB. Снова вам лучше поместить копию в несколько компакт-дисков.

OPC

Второй шаг - это проведение OPC (или создание PDF-файлов для редко используемых документов). Это займет около 500 - 700 месяцев для всего процесса OPC и перевода в Word и HTML при использовании работников на полставки. Это неэффективно, поэтому вам лучше поручить эту работу профессионалам.

Для снижения затрат большинство редко используемых страниц (скажем, 80% или 80.000 с.) можно конвертировать в файлы PDF, а остальные 20.000 в Word, HTML. Файлы PDF заняли бы 4-6 GB, долго загружаемых через Интернет, но стоили бы они $0.2 за страницу у профессиональной компании ($16.000 для 80.000 с.). Если использовать наемный труд для конвертации TIFF в PDF, то это заняло бы 10-20 месяцев работы на полставки.

Поручение другим

Если сохранить пропорцию 80% PDF и 20% HTML, Word, то первая часть стоит около $16.000, а вторая - около $30.000-$40.000, а в целом около $50,000. Если все провести через OPC, это будет стоить $150.000-$200.000.

5
Создание электронной коллекции

Следует учитывать три важные вещи при принятии решения о создании электронной коллекции. Первое, нужно чтобы коллекция была организована и имела определенную структуру. Чем больше содержание, тем разнообразнее должны быть индексы и необходима мощная поисковая система. Для коллекций объемом 3000-5000 страниц очень важно иметь требования, описанные выше. Во-вторых, прежде всего следует учитывать нужды конечного пользователя. Следует четко определить группу людей, для которых создается эта коллекция, и непременно оказывать им консультационные услуги. И наконец, наличие средств определит объем потенциальной работы.

5.1 Методы создания коллекций

Существует много примеров превосходных компакт-дисков, которые созданы по модели веб-страниц. HTML, PDF или документы Word добавляются и связываются с использованием гиперссылок. Навигация сделана простой и привлекательной при помощи гиперсвязей, фреймов, ключевых слов, индексов и т.д. Такие системы работают хорошо вплоть до нескольких тысяч страниц, но с 3000 до 5000 страниц важно иметь хорошо структурированную коллекцию и мощные средства поиска. Именно в этом случае программное обеспечение Greenstone может эффективно помочь.

Программное обеспечение Цифровой библиотеки Greenstone создает структурированную цифровую библиотеку и включает очень мощные поисковые возможности и средства восстановления. До 150000 страниц может быть проиндексировано на отдельном CD-ROM. Каждый CD-ROM может стать сервером Интернета. Greenstone является открытым программным обеспечением с открытой исходной программой и свободно доступно согласно лицензии GNU.

Сопутствующие руководства описывают, как строить коллекции Greenstone. Есть по существу три различных пути их создания:

- Интерфейс библиотекаря
- Коллектор
- Построение из командной строки.

Первый метод - "библиотечний" интерфейс, описанный в руководстве пользователя по Цифровой Библиотеке Гринстоун (Глава 3, «Создание коллекций Гринстоун»). Это - всестороннее диалоговое средство для построения коллекций. С этим, Вы можете строить наборы документов, импортировать или назначать метаданные, и объединять их в коллекцию Гринстоун. Второй метод - "Коллектор" подсистема, описанная в 4 Главе в руководстве пользователя. Это – предидущее средство, которое обеспечивает альтернативный путь построения коллекций с использованием сети или других документов. Ведущее Вас через последовательность диалоговых веб страниц, которые запрашивают информацию. Однако, это не обеспечивает, добавление метаданных к существующим документам, и - поскольку это – веб интерфейс - это не всегда подходящее средство для коллекций, которые требуют больше нескольких минут для построения. Третий метод состоит в том, чтобы управлять программами для построения коллекций непосредственно из командной строки; это находится в руководстве пользователя по Цифровой Библиотеке Гринстоун (Глава 1). Это дает большее количество гибкости для индивидуального использовании программ, экономя на промежуточных результатах, и могут быть желательны для построения коллекций, которые используют очень много времени. Вам будет также полезно прочитать главу 2 Гида руководстве пользователя чтобы использовать все возможности Гринстоун, чтобы строить продвинутые коллекции.

Есть также четвертый метод для создания и редактирования материалов с использованием, программы называемой Органайзером коллекций. Однако, ее функциональные возможности был заменены выше упомянутым интерфейсом библиотекаря. Описанным в документе Использование Органайзера.

5.2 Обучение за 7 шагов и 15 минут

Лучший способ вхождения в существо дела – это посмотреть и прочувствовать Библиотечный интерфейс. Он состоит в том, чтобы фактически создать маленькую тестовую библиотеку. Если Вы располагаете 15-ю минутами, пожалуйста, следуйте этими шагами, и Вы поймете эту программу намного лучше.

До того, как Вы начнете обучение, сначала инсталлируйте Greenstone (см. *Руководство по инсталляции Greenstone*), который включает *Демонстрационную коллекцию* в формате DLS и ее исходные файлы. **Обратите внимание на то, что если Вы хотите иметь возможность добавлять к Вашей коллекции любой из 140**

документов, входящих в коллекцию DLS (вместо всего лишь 11 из этих документов в Демонстрационной коллекции Greenstone), то Вы должны инсталлировать DLS в качестве одного из примеров библиотек Greenstone. Демонстрационная и DLS коллекции будут установлены в *C:\Program Files\gsdl\gsld\collect* в субдиректориях *demo* и *dls* соответственно. Если Вы предварительно установили Greenstone без DLS и хотите инсталлировать ее, то Вы можете перезапустить CD-ROM с Greenstone и добавить эту коллекцию. Нет необходимости вначале деинсталлировать Greenstone.

Мы предлагаем, чтобы Вы отпечатали инструкции, расположенные ниже, и следовали им шаг за шагом:

1. Запустите Библиотечный интерфейс под Windows, выбирая *Greenstone Digital Library* из раздела *Programs* в меню *Start* и в нем *Librarian Interface*. Если Вы используете Unix, то вместо этого напечатайте

   ```
   cd ~/gsdl
   cd gli
   ./gli.sh
   ```

   где *~/gsdl* – директория, содержащая систему Greenstone.

2. Выберите *New* из файлового меню в горизонтальной линейке меню наверху окна, дайте ему название, например, «*My First Collection*» и заполните Ваш адрес электронной почты и краткое описание коллекции. В меню «*Основа этой коллекции*» выберите «*Демонстрационный пример Greenstone*» или «*Подмножество Библиотеки развития*» (результат будет тот же самый, потому что эти две коллекции имеют одинаковую структуру).

3. Чтобы добавить некоторые документы из Демонстрационной коллекции (или коллекции DLS, если она установлена) в Вашу новую коллекцию, щелкните дважды на папке Greenstone с левой стороны, затем дважды - на коллекции, которую Вы хотите выбрать. Документы в ней показаны внизу. Выберите один из них, перетяните его и поместите на панель с правой стороны. Эта панель представляет коллекцию, которую Вы создаете. Выберите несколько документов и перетяните их в нее последовательно один за другим или используя множественный выбор стандартным способом.

4. Добавьте некоторые из Ваших собственных документов,

которые не входят в Демонстрационную коллекцию или DLS-коллекцию. Закройте папку *Greenstone Collections* на левой панели и щелкните 2 раза на папке *Local Filespace*, перейдите к директории, которая содержит некоторые документы (например, маленькие файлы Word или HTML). Перетащите несколько из них в правую панель, чтобы включить их в Вашу коллекцию.

5. Добавьте метаданные к документам в Вашей коллекции. Пока Вы работали под панелью *Gather*, обозначенной символом *Gather* под горизонтальной линейкой меню наверху окна. Щелкните символ *Enrich* около нее. Документы в Вашей коллекции теперь появляются на панели с левой стороны. Щелкните один раз и исследуйте метаданные, связанные с ними, в таблице «*Элемент… Значение*» наверху справа. Используйте нижнюю панель, чтобы изменить индивидуальные значения, выбирая желательный *Элемент* и существующие значения для него из списка или выписывая новое значение в поле около основания. Добавьте метаданные *Названия, Организации, Ключевых слов* к каждому из Ваших собственных документов, которые Вы ввели в коллекцию. После того, как Вы напечатали каждое значение, вы должны щелкнуть *Append,* чтобы добавить это значение к метаданным.

6. Щелкните клавишу *Create,* чтобы покинуть способ *Enrich* и создать Вашу новую коллекцию. Щелкните кнопку *Build Collection* в основании экрана. Во время построения компьютером коллекции Вы будете получать некоторые сведения о том, что он делает.

7. Когда это завершено, щелкните символ *Предварительный просмотр*, чтобы рассмотреть коллекцию с помощью Библиотечного интерфейса. Проверьте *Названия a-z, Организации*, чтобы иметь уверенность, что Ваши документы были включены в коллекцию. При посещении Ваших домашних страниц Greenstone Вы также найдете, что коллекция была установлена как одна из обычных коллекций.