Table of Contents

Критерии качества архива "Науки и Жизни"

ОСНОВНЫМ КРИТЕРИЕМ работы над созданием архива Науки и Жизни для меня является КАЧЕСТВО РЕЗУЛЬТАТА. Под качественным результатом я имею в виду следующее:

- Чёткий и ровный текст и иллюстрации (визуальная составляющая)
- Достоверная цветопередача, проработка деталей в тенях и в светах
- Отсутствие муара на фотографиях
- Одинаковый размер страниц и полей в журнале
- Масштаб страниц 100% соответствующий реальному размеру
- DJVU файлы с текстовым слоем (возможность поиска)
- Наличие интерактивного оглавления в каждом номере
- Отсутствие ошибок в оглавлениях
- Наличие глобального оглавления по всем номерам с удобной навигацией
- Наличие авторского указателя с удобной навигацией
- Наличие тематического рубрикатора с удобной навигацией
- Небольшой размер архива
- Быстрая работа электронного архива
- Удобство пользования архивом

Для меня НЕ ЯВЛЯЮТСЯ ЗНАЧИМЫМИ следующие факторы:

- Необходимость использовать уже наработанные материалы, вне зависимости от их пригодности к созданию КАЧЕСТВЕННОГО архива
- Необходимость задействовать в работе всех сотрудников редакции вне зависимости от их квалификации в данной области
- Политические факторы, то есть учёт того, что кто-то начинает чувствовать себя виноватым за принятые ранее решения.
- Плата за мою работу

Примеры выбора способов создания архива

В выборе способов создания КАЧЕСТВЕННОГО архива я руководствуюсь соображениями МИНИМИЗАЦИИ ВРЕМЕНИ и требуемой КВАЛИФИКАЦИИ работников: в ряде случаев ПРОЩЕ и БЫСТРЕЕ ПЕРЕДЕЛАТЬ материал “с нуля”, чем перерабатывать имеющиеся материалы.

Пример №1

> Чтобы получить качественные материалы из имеющихся JPG файлов надо провести глубокую цветокоррекцию, с каждой страницей свою, и после этого обязательно произвести сильную правку по развороту и обрезке страниц. Оказалось, что в несколько раз быстрее пересканировать материалы с правильными установками.

> Это происходит потому, как имеющиеся JPG файлы были отсканированы в недостаточном разрешении, с автоматической цветокоррекцией, сохранены в формат, с потерями качества и не всегда сохраняющий DPI (JPG), исходные бумажные страницы обрезаны не всегда ровно,

отсканированы с наклоном, не произведено кадрирование, поля сканера серого цвета.

Более того, большое количество JPG сканов принципиально не могут быть исправлены, так как они были сделаны на основе журналов испорченных на этапе обрезки корешков (разброшюровки).

Пример №2

Для того, чтобы получить качественный результат из плохо выведенного PDF требуется провести гораздо больше работы - собрать цельный PDF из разрозненных файлов, проконтролировать, что все страницы соответствуют журнальным (PDF последней версии), отсканировать отсутствующие, либо неправильные страницы, внедрить в DJVU файл текстовый слой (который добавляется автоматически, если исходный PDF имеет нормальные кодировки шрифтов), проконтролировать, что все страницы в номере одинакового размера (проблемы с полями).

В итоге, перевод номеров за 1998-2005 годы потребовал в разы больше работы, чем даже сканирование тех же самых номеров с бумажных экземпляров. И нет гарантии, что полученный электронный архив соответствует бумажному.

Пример №3

Получить качественный результат из уже сделанного оглавления за 1970 - 1997 года тяжелее, чем перенабрать оглавление заново.

Старое оглавление было сделано автоматическим распознаванием в Finereader'e, что привело к появлению большого количества неверно распознанных символов. Большинство фамилий отсутствуют в словарях проверки FineReader'а, так что качество распознавания оглавлений в целом ниже, чем обычного текста (это особенность всех OCR программ).

Вычитывать опечатки и исправлять их получается дольше, чем заново напечатать. Тем более, что люди, занимающиеся введением оглавления в XLS формат - это профессиональные машинистки.

Кроме того, старое оглавление не содержит большого количества информации, требуемой для создания электронного архива - там нет номеров страниц, соответствующих страницам DJVU файлов, нет содержимого рубрик, не проставлены связи между статьями и подвёрстками.

В результате машинистки вынуждены выполнять несвойственную им работу, что приводит к увеличению количества ошибок, по сравнению с оглавлениями, набранными "с нуля".

Практически все переделки требуют гораздо более высокой квалификации, чем проведение работы правильно “с нуля”!

Процесс создания электронного архива "Науки и Жизни"

Исходя из требований качества, снижения трудоёмкости этапов и снижения требований к уровню квалификации участников мной была выработана следующая схема процесса создания электронного архива:

1) Каждый бумажный номер журнала должен проходить следующие ПОСЛЕДОВАТЕЛЬНЫЕ этапы обработки:

- Поступление номера на обработку
- Сканирование в TIFF (на сканере OpticBook - в ч/б, серые и цветные сканы, 300 dpi, правильное кадрирование)
- Проверка полученного материала на качество
- Выборочная цветокоррекция (страницы в градациях серого и цветные)
- Сохранение обработанных TIFF файлов на DVD (TIFF архив)
- Выборочное исправление наклона (шаг можно пропустить)
- Перевод TIFF → DJVU

далее два ПАРАЛЛЕЛЬНЫХ процесса	
Создание оглавления в XLS	Создание текстового слоя (OCR)
Проверка и вычитка оглавления	Разметка активных ссылок на странице “в номере”
Сборка оглавления XLS → HTML	
слияние параллельных процессов	

- Внедрение оглавления в DJVU файл
- Окончание обработки номера

2) Параллельно с этим идёт процесс разработки программного обеспечения для создания глобального оглавления, авторского указателя и тематического рубрикатора.

3) Новые номера (за 2006 год и далее) должны проходить через другую ПОСЛЕДОВАТЕЛЬНУЮ технологическую цепочку:

- Поступление номера на обработку
- Перевод PDF → DJVU
- Создание оглавления в XLS
- Проверка и вычитка оглавления
- Сборка оглавления XLS → HTML
- Внедрение оглавления в DJVU файл
- Окончание обработки номера

Критерии для PDF файлов

PDF файлы должны удовлетворять следующим требованиям:

- Единый PDF Файл должен содержать все страницы журнала в правильном порядке

- Все шрифты целиком внедрены в PDF
- Кодировки внедрённых шрифтов не являются “Custom”
- PDF файл создан без ограничений на печать/копирование и т.д.
- PDF содержит страницу “В номере” с размеченными активными ссылками
- Иллюстрации должны быть сохранены в RGB, так чтобы цвета соответствовали отпечатанному журналу
- Внедрённые картинки должны быть сохранены со следующими параметрами:
    1. цветные и серые: 300 dpi, JPG, максимальное качество
    2. чёрно-белые в исходном разрешении (downsample off), CCITT G4

Качество DJVU

По поводу специальных требований к страницам, заменяющим PDF страницы.

Я пропускал их через Кромсатор, чтобы выровнять наклон. После этого пропускал через кромсатор, чтобы выровнять размер страниц - размер брал из DJVU файлов, сделанных из PDF - в свойствах страницы указан размер с точностью до пикселя.

Была ещё одна тонкость - надо было принудительно добавлять поля снизу - special gap, дабы выравнивание текста на получающейся странице было такоеже, как в оригинальной вёрстке.

Слушай, а почему в 1992_01 чёрно-белые страницы закодированы как серые? Где-то лажануло…

И в номере 1992_02 фон как-то странно удалён - остался серый фон и видны границы страницы…

например 1992_02 стр 6, 19, 28, 30, 38…. Что-то тут не так.

стр 57 и так далее.

Ты чем в DJVU конвертировал? Это к вопросу, что ч/б страницы закодированы как цветный?

Кстати, обложки отсканированы с обрезкой корешка, что не есть гуд.

Практически все номера сделаны с подобными глюками - чёрно белые страницы закодированы как полноцветные, фон убран не до конца, обложки избыточно обрезаны.

Объясни, пожалуйста, как ты делал эти файлы?

хм… так как мы с тобой договаривались, так и делал: 1. проход бетчем фотошопа (цветокоррекция) 2. ручной просмотр всех страниц, выборочная цветокоррекция и коррекция наклона 3. компрессия в дежавю с использованием твоего скрипта где мог быть глюк…?

бетч из пункта 1 как базовый я брал твой с сайта

У тебя после прохода фотошопом должны были получиться TIFF ч/б страницы черно белыми (двуцветными, а не RGB). И серые серыми (256 цветов).

На пункте 2 ты явно пропустил те страницы, что я тебе из первого номера накидал как страницы с некачественно удалённым фоном.

Скрипт для фотошопа требует настройки - там надо подстраивать один из этапов levels - я об этом писал.

Скорее всего ты на этапе фотошопа изменил цветовое кодирование в RGB для всех страниц.

И на счёт обложек - либо ты их подрезал наряду со всеми внутренними страницами, либо они были отсканены некачественно. Соответственно надо их либо по-новому отсканить, либо по-новому цветокорректировать из необрезанного бэкапа.