# Подвиг в цифре

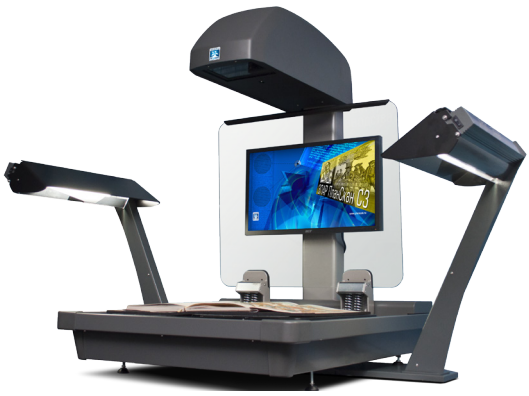
Мир ПК
*Анатолий Воронин*
№ 11, 2013 г.

Возможно, вы уже знаете о проектах ОБД «Мемориал» и «Подвиг Народа», на которых выложены документы из военных архивов. Сами проекты принадлежат Министерству обороны Российской Федерации, а их техническим воплощением занималась компания ЭЛАР. Ее задачей стала оцифровка миллионов страниц документов из архивов и создание на их основе базы данных, по которой можно было бы вести поиск необходимой информации.

Первый этап работы над этими базами закончен, сейчас на очереди их расширение. В настоящее время в проекте «Подвиг Народа» перечислены почти все Герои Советского Союза и орденоносцы, а также обладатели медалей «За Отвагу» и «За боевые заслуги». Награжденные другими медалями, например, «За оборону Москвы» (Ленинграда, Одессы) или «За взятие Берлина» (Праги, Вены), оказались неучтенными. Кроме того, в работе использовались только документы и картотека ЦАМО (Центрального Архива Министерства Обороны), а информация о награждениях и награжденных находится и в других архивах. Например, в ЦАМО нет данных о подвиге Виктора Талалихина, совершившего ночной таран в подмосковном небе. Весь этот массив информации должен быть добавлен к 70-летию Победы.

Впрочем, некоторые категории награжденных останутся по-прежнему секретными — информация о наградах бойцов НКВД исключена из публичного доступа.

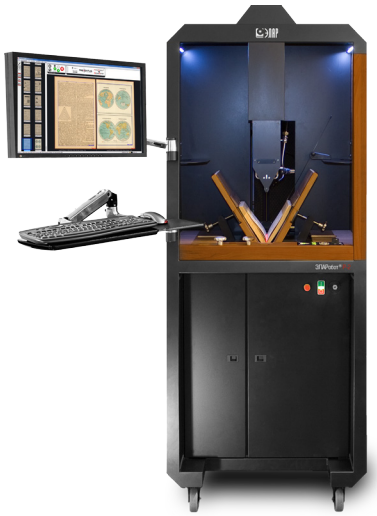
Для оцифровки всего массива документов (а это 19 млн листов) были разработаны специальные сканеры. Планетарный сканер сильно отличается от более привычного нам планшетного. Его сканирующая головка находится высоко над документом, фактически фотографируя его сверху. В ней может быть установлена либо ска-



нирующая линейка, либо, как в нашем случае, матрицы. В сканерах, использовавшихся при обработке документов в ЦАМО, установлено по три 80-Мпикс матрицы RGB. Они захватывают область чуть больше формата А3, что позволяет сканировать за один цикл два стандартных листа размера А4.

Сканирование бесконтактное, хотя на практике листы все же прижимают стеклом. Поверхность сканирования может «переламываться», образуя книжную «колыбель», — дела можно сканировать без расшивки, что в ряде случаев принципиально. Кстати, для книг существует особый сканер, который сам нежно листает страницы.

И вот коробка документов просканирована, а дальше начинается самое интересное и сложное. Сам по себе образ документа практически бесполезен без распознания содержащейся в нем информации. Казалось бы, созданы замечательные программные комплексы для машинного распознавания документов, и извлечение информа-

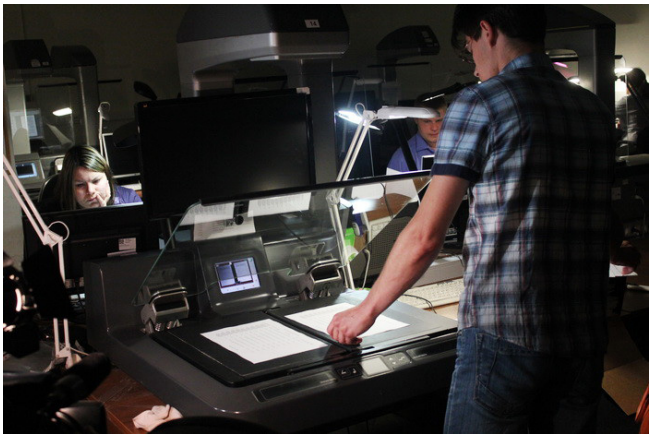


ции из сканов вопрос лишь машинного времени. Но, как оказалось, при распознавании документов 70-летней давности, даже напечатанных на машинке, доля ошибок колеблется вокруг 50%-й отметки. На их исправление оператор тратит больше времени, чем на ввод той же информации вручную. Вот такая техническая загогулина. А если учесть, что значительная часть документов вообще написана от руки, становится очевидным, что без человека опять не обойтись.

В настоящее время проект ОБД «Мемориал» содержит 29 млн записей, а «Подвиг народа» — 12 млн. Цифра 29 млн не означает общее количество погибших, зачастую на одного бойца может приходиться по нескольку записей, на каждую из которых уходит около 3 мин рабочего времени оператора. В итоге счет идет на миллионы человеко-часов.

Для решения столь сложной проблемы разработчикам пришлось призвать целую армию операторов-«надомников» — 5 тыс. человек. Причем не только из Москвы, но и из российских регионов. Для того чтобы минимизировать ошибки, каждый документ дважды распознавался разными операторами. Если их результат совпадал (это легко проверяется машиной), то данные заносились в базу. В противном случае документы шли по новому кругу или передавались эксперту, который лучше разбирается в почерках.

Более того, каждая запись содержит восемь — десять полей. Операторам выдавались образы лишь одного поля (ячейки таблицы), без возможности посмотреть весь документ. Это не только препятствовало утечке информации и способствовало более качественному распознаванию, но и щадило нервную систему операторов. Ведь читать подряд «Донесения о безвозвратных потерях» — очень



тяжелая психологическая нагрузка. А вот вычитка номеров частей или мест призыва совсем не угнетает.

И это еще не вся выгода от разбивки по полям. Объем данных столь велик, что разработчики на первом этапе сосредоточились на внесении только основных данных, что позволило ускорить и удешевить создание системы. Информация о местах первичного захоронения, а также об адресе родственников погибших бойцов будет распознана позже. Пока ее можно прочесть только на электронных копиях документов.

Сколько всего страниц придется отсканировать в рамках совершенствования проекта, разработчики пока не сообщают, но, похоже, их операторам придется не менее года разбирать карандашные записи, сделанные в окопах и блиндажах.

И вот документ отсканирован, распознан — и что же дальше? А дальше информация заливается на серверы, расположенные на площадке «Ростелекома». Это мощный провайдер, однако и его ресурсов не всегда хватает при пиковых нагрузках, которые традиционно приходятся на начало мая. В нынешнем году к ресурсу за праздничные дни обратились 1 млн раз, причем 9 мая число их число достигло 180 тыс. человек, что привело к временным перебоям.

По словам разработчиков, база построена по образцу карт Google — образы документов состоят из «тайлов», отдельных квадратиков, складывающихся в мозаику. Это позволяет существенно уменьшить нагрузку на сервер при отдаче документа.

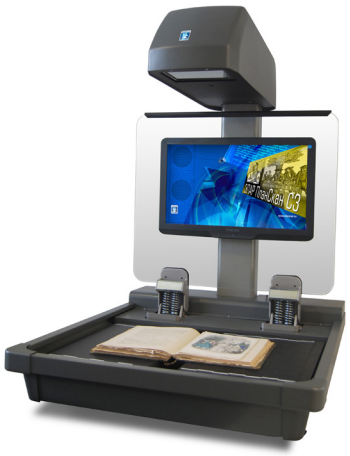
Одновременно такое решение стало и своеобразной защитой от копирования документов. Для того чтобы собрать оригинал, надо скачать все «тайлы» при 100%-м увеличении и собрать из них целый лист. Разработчики утверждают, что таково было требование заказчика — Минобороны России, для того чтобы осложнить жизнь «черным копателям». Впрочем, для документов о погибших или приказов о награждениях сделано исключение — их можно сохранить без проблем.

Обращение к базам ведется через поисковые запросы. Причем, в отличие от «Яндекса» или Google, чем запрос менее конкретен, тем успех более реален. Дело в том, что в исходных документах очень много ошибок, зачастую они бывают неполные. Например, не указаны отчество или год рождения, что приводит к непониманию запроса поисковой машиной. Наилучший результат получается при вводе минимального количества слов. Лучше вводить минимум информации — как правило, только фамилию и имя. И уже потом добавлять данные в другие поля.

Армейские писари не отличались большой грамотностью и, записывая со слуха, зачастую путали буквы. А в окончании фамилии или отчества может быть форменный кошмар. Это выправляется подстройкой условий в расширенном поиске. Вместо заданного по умолчанию «Точная фраза» можно использовать «С начала поля» или «Полнотекстовый поиск».

Надо понимать, что электронные базы ОБД «Мемориал» и «Подвиг Народа» — это отображения бумажных документов. Если ошибка содержится в бумаге — она перекочует в электронную опись. И данная проблема пока не разрешена. Предложения и замечания по исправлениям принимаются, но не всегда отражаются в базе. Исправляют только очевидные ошибки, причем определение степени очевидности разработчики оставляют за собой. Конечно, они согласны с тем, что база должна совершенствоваться, но для такой работы нужны уже не операторы-надомники, а люди, способные более глубоко анализировать информацию, находить связи, не лежащие на поверхности. Так что с учетом количества записей и необходимости больших трудозатрат при анализе подобная работа растянется на долгие годы.

На помощь может прийти новый проект, который вызревает в недрах Минобороны России и компании ЭЛАР. Суть его — в объединении баз существующих проектов с добавлением в них географической составляющей. Идея довольно простая: каждый человек или воинское подразделение в каждый момент времени находится лишь



в одной географической точке. Таким образом, все происходившие события можно привязать к географической карте. Казалось бы, легко и просто, но дальше начинаются многочисленные «но».

В первую очередь, необходимо создать векторную географическую карту Европейской части СССР по состоянию на 1941–1945 г. В настоящее время многие населенные пункты уже не существуют, их координаты и границы неизвестны. Привязка к карте мест расположения частей и подразделений — также очень непростой процесс, который возможно реализовать лишь вручную. Расположение частей и подразделений поможет локализовать места первичных захоронений. Обычно они привязывались к конкретным населенным пунктам, которые находились рядом с местами дислокации частей. Это, в свою очередь, позволит более точно заполнить места первичных захоронений, информация о которых не распознавалась.