УДК 025.133:004:070

Брайан Бенилоус
«Ист Вью Информейшн Сервисиз», директор газетных ресурсов, США, Миннеаполис, e-mail: bryan.benilous@eastview.com
Барбара Крупа
Стэнфордский университет, управляющий проектом оцифровки газет, США, Стэнфорд, e-mail: krupa@stanford.edu
Джеймс Саймон
«Центр научных библиотек», вице-президент по фондам и услугам, США, Чикаго, e-mail: simon@crl.edu
Фредерик Зарндт
Компания «Диджитэл Дивайд Дэйта», консультант, США, Нью-Йорк, e-mail: frederick.zarndt@digitaldividedata.com

Архив мировой прессы «Ист Вью»: уроки программы оцифровки газетных материалов

Аннотация. В статье освещены особенности и возможности партнёрства как успешной модели поддержки процессов оцифровки и распространения массивных газетных фондов, а также уроки, полученные в ходе реализации описываемой программы. Среди главных проблем названы: логистика; «облачные» интерфейсы; издатели и «копирайт»; работа с языками, использующими нелатинские шрифты; расширение доступности ресурсов посредством консорциумов и фондов открытого доступа.
Ключевые слова: газеты; оцифровка газет; проекты; открытый доступ; консорциумы.

Bryan Benilous
East View Information Services, director Newspaper Products, USA, Minneapolis, e-mail: bryan.benilous@eastview.com
Barbara Krupa
Stanford University, Newspaper Digitization Project manager, USA, Stanford, e-mail: krupa@stanford.edu
James Simon
Center for Research Libraries, vice president of Collections and Services, USA, Chicago, e-mail: simon@crl.edu
Frederick Zarndt
Digital Divide Data, consultant, USA, New York, e-mail: frederick.zarndt@digitaldividedata.com

«East View» Global Press Archive: lessons learned from the massive newspaper digitization program

Abstract. This paper highlights both partnerships as successful models for supporting the digitization and distribution of massive newspaper collections, and highlights key lessons learned from the program. Topics include: logistical challenges; development of cloud-based workflows; publishers and copyright; challenges of working with languages in non-Roman scripts, and broad accessibility through consortia and Open Access collections.
Keywords: newspapers; newspaper digitization; digitization projects; Open Access; consortia.

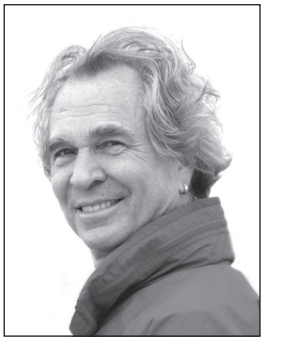
Б. Бенилоус

Б. Крупа

Дж. Саймон

В 2016 г. компания «Ист Вью Информейшн Сервисиз» (ИВИС) совместно с библиотеками Стэнфордского университета и библиотекой и архивами Гуверовского института начала оцифровку для обеспечения доступа к 25 млн газетных страниц, напечатанных в основном на китайском, японском и арабском языках, к которым были частично добавлены и газеты на кириллице. Такое сотрудничество послужило стимулом для того, чтобы значительно расширить свою собственную программу оцифровки газет, известную под названием «Архив мировой прессы Ист Вью». В 2019 г. ИВИС также привлекла к партнёрству и Центр научных библиотек (ЦНБ), чтобы сделать доступными ещё 4,5 млн газетных страниц с помощью программ открытого доступа и фондов, базирующихся на принципах консорциума. «Архив мировой прессы» должен обеспечить широкий и открытый доступ к газетному контенту на редких языках со всего мира, а также поддержать развитие цифровых гуманитарных исследований.

Библиотеки Стэнфордского университета и библиотека и архивы Гуверовского института Фонды документов Гуверовского института[1] начали формироваться в 1919 г. изданиями по тематике «Война, революция и мир». Кураторы, работавшие в разных точках земного шара, стремились собирать подобного рода документы, уделяя особое внимание газетам, которые оперативно отражают историю и мнения людей. В итоге Гуверовский институт собрал газеты из 125 стран, и их фонд в сумме превышает 25 млн страниц. В фонд попали не только главные ежедневные газеты, но и редкие издания, которые не были представлены ни в одной библиотеке западных стран.



Ф. Зарндт

В 2001–2002 гг. библиотеки Стэнфордского университета (БСУ) получили около 60% вышеозначенных фондов Гуверовского института, в том числе относящихся к Восточной Азии. Печатные газетные издания составляли львиную

[1] Учреждение названо в честь 31-го президента США Герберта Гувера, который был выпускником (в 1895 г.) Стэнфордского университета и, по сложившейся традиции, финансировал там строительство здания для мемориальной библиотеки и архива. — *Примеч. переводчика.*

долю в этом перемещении. Около 2700 газетных названий из многих стран и на многих языках были перевезены из «Башни Гувера» в Стэнфорде в главное хранилище в Ньюарке (штат Калифорния), которое расположено в 17 милях к востоку от главного кампуса. Персонал выводит на экраны газеты по запросам пользователей два раза в неделю и обеспечивает доступ к ним через «Зелёную библиотеку» или Восточноазиатскую библиотеку (обе входят в комплекс БСУ)[2].

Газетный фонд охватывает Восточную Азию, Средний Восток (включая Турцию), Африку, обе Америки, Западную и Восточную Европу, Новую Зеландию и Австралию. Большинство изданий датируется со второй половины XX в. Имеется некоторое количество редких изданий 1940–1950-х гг. из Китая, а также почти исчезнувших изданий из Африки. Есть издания, имеющиеся только в фонде БСУ. Некоторые газеты, в особенности африканские, печатались недолгое время, а другие насчитывают многие десятилетия своей истории и, соответственно, хранения. Газеты обычно хранятся в пачках или архивных коробках. Большинство их в хорошем состоянии.

БСУ получили фонды со многими информационными пробелами. Некоторые названия не были отражены в онлайновом библиотечном каталоге, а другие, будучи включёнными, содержали только минимальную библиографическую информацию, например, каталожный индекс, заглавие и название фонда. Из-за этого поиск точной записи в OCLC и верификация актуальных фондов под вопросом. У некоторых газетных названий неправильно указано место хранения, имеется неточная датировка, у других неправильное орфографическое обозначение, есть пропущенные наименования и потерянные документы (утрата произошла, по всей вероятности, при исключении из фонда после микрофильмирования без нужного обновления записи). В 2012–2014 гг. сотрудники Восточноазиатской библиотеки предприняли чистку онлайнового каталога в хранилище Ньюарка. Газеты на китайском и японском языках были выделены из общей массы газет, заново упакованы и переписаны. Также была проведена верификация соотношения БСУ и Гуверовских фондов на микрофильмах и газетных печатных фондов для всех географических регионов.

В 2016 г. БСУ с библиотекой и архивами Гуверовского института решили оцифровать все газеты из хранилища Ньюарка, чтобы сделать их полностью доступными для университетского сообщества в режиме 24/7. Сканирование и оцифровка начались в октябре при сотрудничестве с «Ист Вью». Руководителем этого процесса была назначена Барбара Крупа, которая должна была обеспечить координацию работы всех участников в лице «Ист Вью», различных подразделений БСУ и персонала Гуверовского института.

История деятельности Центра научных библиотек в сфере сохранения газетных фондов и создания их общей коллекции

Центр научных библиотек (ЦНБ) был создан в 1949 г. как Межбиблиотечный центр Среднего Запада США: уже тогда осознавалась важность сохранения международных фондов газет. Общая их коллекция сначала сформировалась как репозитарный фонд из поступлений от библиотек — членов, которые желали освободить у себя площади для хранения новых поступлений. По мере роста читательского спроса на газеты из са-

[2] Библиотеки крупных университетов США, как правило, состоят из автономных, территориально разбросанных подразделений со своими фондами и коллекциями для обслуживания информационных запросов студентов и преподавателей по определённым темам и дисциплинам. — *Примеч. переводчика.*

мых разных регионов ЦНБ начал подписываться на микроформные издания (а также сам производил микрофильмирование) 57 зарубежных печатных газет. Этот процесс начался в 1952 г., а уже в 1956-м ЦНБ стал штаб-квартирой масштабного проекта микрофильмирования зарубежных газет. В рамках проекта была обеспечена закупка 100 названий зарубежных газет на микрофильмах, которые затем должны были предоставляться заинтересованным организациям во временное пользование или для заказа копий для собственного хранения.

ЦНБ также действует как общая научная платформа, которую должны обеспечить своим пользователям различные библиотеки для международных и региональных исследований (главным информационным ресурсом выступают газеты). И именно ЦНБ обладает здесь огромным опытом поддержки, создания и хранения общего, хотя и распределённого библиотечного фонда, в том числе редких изданий, а также оперативного предоставления конкретным организациям в открытый доступ нужных ресурсов. Уже в 1963 г. с основанием и запуском проекта кооперации микроформных изданий «Африканы» ЦНБ стал администратором этой программы. Вслед за этим начались и другие проекты аналогичной направленности. В 1967 г. был запущен микроформный проект для Южной Азии, в 1970 г. — для Юго-Восточной Азии, в 1975 г. для Латинской Америки, в 1987 г. для Среднего Востока, в 1995 г. — для славянских стран и Восточной Европы. Недавно из названий этих шести проектов было убрано слово «микроформы» — его заменили термином «материалы». Такая замена отражает изменения в составе библиотечных фондов газетных изданий, в которых уже преобладают цифровые носители. Все проекты по-прежнему нацелены на приобретение, хранение и использование редких и труднонаходимых научных информационных материалов из конкретных (обозначенных в названиях самих программ) регионов мира или/и о них. Каждый проект финансируется ежегодными взносами институционных участников для приобретения и перезаписи актуальных и важных для пользователей коллекций газетных изданий. ЦНБ хранит их, обеспечивая возможность пользования этими фондами на условиях межбиблиотечного абонемента или цифрового удалённого доступа.

Кроме реализации своих приоритетов в сфере газет и мировых информационных ресурсов, ЦНБ создал коллекцию международных новостных ресурсов для научных исследований и целей образования. Каталог этих новостных ресурсов насчитывает свыше 18 тыс. записей начиная с середины XVIII в. Что касается газетного фонда, то его 25% — это общенациональные газеты США, выпускавшиеся и распространявшиеся во всех штатах и территориях с колониального периода до настоящего времени. Иностранных газет в фонде ЦНБ насчитывается около 10 тыс. названий, также выходивших с середины XVIII в. Основное место в этом фонде занимают газеты из Европы и Великобритании, Латинской Америки, Африки, Китая, России, Южной Азии.

Библиотеки Северной Америки заслужили славу, обеспечивая быстрый электронный доступ к новостному контенту. Однако огромная часть аналогичной по содержанию новостной информации в Южном полушарии по-прежнему распространяется в печатном виде. Пока что сделано не так уж много для масштабной оцифровки этих ресурсов. С 2006 г. ЦНБ пытается ликвидировать этот пробел, сканируя различные материалы по конкретным запросам своих институциональных членов. К 2019 г. оцифровано более 300 тыс. разных выпусков из 900 газетных названий, в том числе тех, которые были ранее переведены на микрофильмы.

ЦНБ уделяет особое внимание выработке и соблюдению стратегии широкомасштабной оцифровки новостной информации. В 2008 г. ЦНБ совместно с «Ридекс» (подразделение «НьюсБанка») запустил в действие проект «Мировой газетный архив», который подразумевает многоэтапную систематическую оцифровку и архивирование новостных газет всего мира и других новостных материалов из фондов североамериканских научных библиотек. К настоящему времени оцифровано уже более 3,5 млн страниц текста примерно из 0,5 млн выпусков 400 газет, выпущенных в Латинской Америке, Африке и Южной Азии. Как уже говорилось, существуют и давно реализуются проекты оцифровки газетных массивов, специализированные по конкретным мировым регионам и нацеленные также на выявление и сохранение новостных ресурсов из главных газет XIX и XX вв.

Цели сотрудничества ЦНБ с «Ист Вью»

Несмотря на достигнутые успехи, для новостных записей культурного или исторического значения по-прежнему существует угроза исчезновения. Оцифровка новостей из Европы, Северной Америки и других развитых регионов продолжает успешно развиваться, особенно для изданий конца XIX — начала XX в., что делает доступ к новостным изданиям Северного полушария всё более свободным и удобным. Однако новостной контент из изданий Южного полушария в онлайновом режиме менее доступен, а применительно к изданиям со шрифтами нелатинского происхождения — особенно затруднён и предоставляется пользователям очень фрагментарно.

Кроме того, этот доступ ограничен проблемами авторских и издательских прав, что блокирует онлайновый доступ к большому объёму новостного контента из изданий второй половины XX в. По мере развития оцифровки всё чаще члены ЦНБ выражают желание расширять доступ к современным новостным ресурсам, чтобы компенсировать разрыв между юридическими ограничениями в сфере копирайта, введёнными ещё в начале XX в., и полнотекстовыми базами данных различных статей, которые стремительно развиваются сегодня. Участники ЦНБ и другие крупные библиотеки агрегировали уникальные новостные коллекции, представляющие разные точки зрения и языки, на которых эти мировоззрения выражены. Разблокирование таких баз данных открывает перспективы для учёных и исследователей в сферах языкознания, текстологии и поиска текстов для цифровых гуманитарных наук.

В 2016 г. была провозглашена политика, согласно которой пользователям должны быть доступны без ограничений все цифровые ресурсы, которые созданы на базе материалов, уже переведённых в общественное пользование либо ещё защищённых копирайтом, который был выкуплен ЦНБ. Возрастающее стремление библиотек — членов ЦНБ создавать уникальные цифровые коллекции для широкого общественного доступа стало важным фактором в развитии новых моделей партнёрства и сотрудничества. Уже в начале переговоров с «Ист Вью» по поводу его программы «Архив мировой прессы» члены ЦНБ выражали максимальную заинтересованность именно в выработке условий обеспечения широкого общественного доступа к таким ресурсам.

Роль компании «Диджитэл Дивайд Дейта» в оцифровке газет и в сфере благотворительности

Эта компания под сокращённым названием DDD была основана в 2001 г. в Пномпене (Камбоджа) десятью физическими лицами. Сегодня штат DDD превышает 3 тыс. человек, работающих в Азии, Африке, на Среднем Востоке

и в Северной Америке. Когда Джереми Хокенстейн, основатель компании, проезжал по Камбодже, его поразила смешанная картина нищеты и прогрессивных изменений в стране, начавшей восстанавливаться после падения режима «красных кхмеров». В стране уже работали школы компьютерной грамотности и профессиональной подготовки в сфере информационных технологий для молодёжи, но не было рабочих мест для выпускников этих заведений. С учётом возможности перенести в Восточную Азию уже опробованный в Индии опыт аутсорсинга информационного бизнеса, DDD открыла небольшой офис в Пномпене.

Сегодня DDD поставляет обработанный контент сотням зарубежных и местных пользователей, используя свои операционные центры в Пномпене и лаосском Вьентьяне, кенийском Найроби и арабском Секторе Газа, филиппинской Маниле и американском штате Вирджиния. DDD является крупнейшим работодателем в сфере высоких технологий для Камбоджи и Лаоса. Уникальная модель DDD «Импэкт Соурсинг» объединила усилия многих молодых профессионалов и вызволила из нищеты сотни их семей. Эта модель является фрагментом системы аутсорсинга информационного бизнеса, которую в данном регионе внедрила именно DDD, доказав её эффективность как экономически оправданного подхода к решению проблемы нищеты. Этот подход обеспечивает высокое качество контента и информационных услуг, а также формирует рабочий персонал, вполне конкурентоспособный в мировом масштабе.

В Камбодже, Лаосе, Кении и Секторе Газа главной проблемой для образованной молодёжи является отсутствие требующихся ей рабочих мест на традиционном рынке труда. А малообразованная и не имеющая профессиональной подготовки молодёжь в этих странах сталкивается с безработицей либо вынужденно связывается с теневым бизнесом (а то и с криминалом). В любом случае заработки у молодых людей обычно невелики, при этом какая-либо охрана труда и гарантия социальной защиты отсутствуют.

Вышеописанная модель социального воздействия даёт молодому специалисту возможность стать официальным сотрудником DDD и учиться там коммерческому аутсорсингу информационного бизнеса, включая создание требуемого цифрового контента, подготовку данных, машинное обучение и услуги облачных технологий для пользователей во всём мире. Через DDD можно приобрести современную профессиональную подготовку и компетенции, а также займы и стипендии для получения высшего образования. Такой подход основан на целостной программе DDD по задействованию профессиональной подготовки, максимальной занятости молодёжи, обеспечению её высшего образования, чтобы каждый участник смог самостоятельно определить нужные жизненные цели и достичь их. За последние 18 лет такое обучение в рамках программы и поддержку получили более 6 тыс. безработных молодых людей, и они заработали за это время более 350 млн долларов.

DDD получила широкое признание как лидер в сфере развития человеческого капитала. В 2008 г. она получила престижную награду Международного некоммерческого фонда Джеффа Сколла («Сколл Эворд»), а Международная ассоциация профессионалов аутсорсинга (IAOP) включила DDD как «восходящую звезду» в список своих 100 общемировых лидеров в 2015 и 2016 гг. Кроме того, DDD постоянно входит в список 100 неправительственных организаций — мировых лидеров информационного бизнеса, а также получила грант от компании «Гугл» за инновации в этой сфере.

DDD считается лидером в сфере аутсорсинга информационного обслу-

живания процессов оцифровки и сохранения фондов библиотек, частных и общественных архивов, содержащих редкие коллекции, а также научных университетов и музеев во всём мире. Эти услуги расширяют информационную ценность и доступ к различным данным, документам, публикациям и архивам — в особенности при переводе их в онлайновый режим для мобильных устройств считывания или в информационную систему. Эта компания достигла ежемесячной производительности (оцифровка, создание метаданных), измеряющейся более чем в 0,5 млн страниц текста.

DDD зарегистрирована в США как некоммерческая корпорация, призванная создавать надёжные и стабильные рабочие места и гарантировать получение образования в развивающихся странах.

DDD владеет коммерческими компаниями в Камбодже, Лаосе, Кении и США, а также тесно сотрудничает с аналогично нацеленными на развитие человеческого капитала компаниями в Секторе Газа и в Маниле. Деятельностью DDD руководит совет директоров из 12 человек. Прибыль DDD в 2018 финансовом году составила более 18 млн долларов.

Основной штат руководящих сотрудников DDD накопил уже более чем 10-летний опыт экспертизы и проведения сканирования, тегирования метаданных, оцифровки. До работы в DDD эти люди уже имели профессиональный опыт, полученный в фирмах США, Индии, Сингапура и Филиппин.

Главный офис DDD расположен в Пномпене в шестиэтажном здании (и с уже семилетней историей), где размещено 500 рабочих мест. При этом четыре этажа служат именно для производственных целей. Полезное пространство может быть расширено до 725 рабочих мест.

Оцифровка газет была и остаётся одной из основных сфер деятельности DDD. Также эта корпорация расширяет сотрудничество с музеями и библиотеками в Африке и Азии для сохранения их фондов и коллекций в результате использования новейших методов и технологий оцифровки, хранения и обеспечения доступа к требуемому контенту. В ходе реализации различных проектов новые и молодые сотрудники DDD не только развивают свои навыки и компетенции, что укрепляет их позиции на рынке труда, но и повышают культурный уровень, способствуя при этом сохранению исторического наследия страны для будущих поколений.

Среди ключевых проектов DDD — оцифровка находившихся под угрозой разрушения коллекций в национальных музеях Кении и в камбоджийском Музее геноцида «Туол Сленг», где собраны многочисленные письменные и фотодокументы периода «красных кхмеров».

DDD поддерживает деятельность Американской библиотечной ассоциации (ALA) и Международной федерации библиотечных ассоциаций и учреждений (IFLA), как и других ассоциаций и организаций схожего профиля. Сотрудники DDD регулярно и активно участвуют в конференциях ALA и IFLA.

Трансформация газетной программы «Ист Вью» в «Архив мировой прессы»

Уже 30 лет ИВИС работает над смычкой издательств и библиотек, а также Запада и Востока для ввода в оборот «малодоступной информации из необычных мест» для поддержки научных исследований. При этом ИВИС стремится к достижению баланса разных точек зрения на научную проблему, в освещении которой обычно доминируют информационные ресурсы Запада.

Если говорить об оцифровке и архивировании газетных изданий, то подавляющее большинство западных агрегаторов контента фокусируется на английском и западноевропейских языках, используемых для своей продукции ведущими издателями. Первые два проек-

та оцифровки газет были связаны с лондонской «Таймс» и американской «Нью-Йорк Таймс». Обе газеты, безусловно, очень важные и авторитетные, тем не менее, представляют читателю лишь собственную, «западную» точку зрения на мировые события. Последующие программы оцифровки (как коммерческие, так и для обеспечения открытого доступа) продолжали фокусироваться на западноевропейских языках, причём не только по причине большей лёгкости обработки таких ресурсов, но и вследствие явного преобладания запросов пользователей и обеспечения надёжного финансирования со стороны заинтересованных библиотек Северной Америки и Европы. Даже сегодня в крупнейших коммерческих программах оцифровки газет очень мало контента на языках с «нелатинской» письменностью. Современные коммерческие ресурсы Восточной Азии фокусируются на англоязычных изданиях из Китая, Кореи и Индии, каковыми часто являются англоязычные региональные газеты либо западные «миссионерские» газеты прошлых лет, а эти источники по-прежнему отражают западное мировоззрение и западную точку зрения. Поэтому «Ист Вью» осознал важность и необходимость создания контента и его адекватного отражения применительно именно к Востоку, а особенно к Восточной Европе, России и бывшим республикам СССР, к Восточной Азии и Среднему Востоку.

Первой оцифрованной для электронного архивирования газетой была «Правда» — главная официальная газета СССР, остававшаяся лидером и в тогдашней России. Затем «Ист Вью» продолжила расширение проектов оцифровки изданий на кириллице из России и Восточной Европы, используя «Известия» и «Литературную газету», при этом всячески развивала распространение оцифрованного контента от своих партнёров из Восточной Азии, например «Jiefanjun Bao» («PLA Дейли»). Также политика «Ист Вью» позволяет учитывать в научных исследованиях все точки зрения на «холодную войну», период разрядки международной напряжённости и другие важнейшие явления и события общемирового и регионального значения.

В 2016 г. «Ист Вью» в содружестве с БСУ и библиотекой и архивами Гуверовского института разработали процесс оцифровки для целой и автономной коллекции газет, собранной Гуверовским институтом в течение XX в.: из 125 стран и объёмом более 25 млн страниц. Этот совместный проект рассчитан на 10 лет сотрудничества, в том числе для выполнения работ на предварительном этапе идентификации уже имеющихся цифровых копий, а затем для оцифровки и коммерческой реализации создаваемых в рамках проекта электронных ресурсов. Внушительная стоимость этих операций для означенного объёма коллекций, а также факт наличия действующего «копирайта» применительно к большому количеству оцифровываемых материалов выдвигают коммерческий аспект на первый план, поскольку для работы необходимо выкупать права на контент у издателей и пр.

Полученный массив газетных материалов позволит «Ист Вью» пересмотреть свою программу оцифровки для выполнения миссии крупнейшего в мире оцифровщика газетного печатного текста. Эта программа была забрендирована под названием «Архив мировой прессы Ист Вью». Основываясь на накопленном опыте оцифровки газет (около 20 названий) ещё до сотрудничества со Стэнфордом и Гуверовским институтом, «Ист Вью» резко ускорила выполнение оцифровки таких газет, как «Москоу Ньюс», «Рафу Шимпо», «Гудок» и «Новое русское слово», которые хранились в фондах партнёров. Одновременно с ростом объёма и разнообразия цифруемого контента она стала разви-

вать дополнительные партнёрские связи, чтобы не только снизить стоимость технологических процессов, но и обеспечить широкий и свободный доступ к новым цифровым ресурсам.

Почти сразу после провозглашения партнёрства со Стэнфордом и Гуверовским институтом «Ист Вью» начала тесно сотрудничать с ЦНБ для дальнейшего продвижения своей амбициозной программы. Результатами этого сотрудничества стали экономически обусловленная доступность материалов на микроплёнке, которые создавались десятилетиями по инициативе именно ЦНБ входящими в его состав библиотеками; широкое распространение оцифрованного контента в научном сообществе при помощи систем открытого доступа и на базе спонсорской поддержки общих фондов консорциумов библиотек; финансирование оцифровки редких материалов при совместном участии всех библиотек ЦНБ, что в противном случае было бы нереализуемо из-за недостаточности рыночного спроса на информационные продукты подобного рода. При этом Стэнфорд и ЦНБ получают все цифровые копии в различных форматах, включая METS/ALTO XML, JPEG2000, PDF с функцией поиска, а также TIFF. В итоге обеспечена долгосрочная цифровая сохранность всего материала из различных хранилищ по США, включая Калифорнию, Иллинойс и Мичиган.

Если сотрудничество со Стэнфордом рассчитано на 10-летнюю оцифровку огромного газетного фонда, то сотрудничество с ЦНБ нацелено на узкий массив из 4,5 млн страниц, которые должны быть оцифрованы за три года.

Этот проект рассматривается как пилотный и с перспективой расширения такого союза для оцифровки не только архива Стэнфорда, но и всего микрофильмного фонда ЦНБ и входящих в его состав библиотек. Также проект нацелен на максимальную производительность, но в любое время возможно её гибкое регулирование для каждого отдельного года.

ЦНБ и «Ист Вью» учитывают следующие пожелания своих членов и клиентов из научного сообщества: создавать возможно больший объём контента для последующего свободного общемирового доступа; избегать принудительных закупок получаемого контента со стороны организаций — членов данного альянса и вводить расценки соразмерно масштабам деятельности и финансовым возможностям конкретных организаций; обеспечивать для ЦНБ и его членов активное участие в выборе контента для оцифровки; особое внимание уделять долгосрочному хранению получаемого цифрового контента и его использованию в цифровых гуманитарных исследованиях.

Описываемая программа была рассчитана на оцифровку девяти коллекций в течение трёх лет, а каждая коллекция имела объём около 500 тыс. страниц текста. Причём члены ЦНБ финансировали эту инициативу в интересах общемирового научного сообщества. Насколько это было возможно, оцифрованный контент (6 из 9 коллекций — более 3 млн страниц) переводится затем в открытый доступ. Оставшиеся пока закрытыми три коллекции содержат защищённый «копирайтами» материал, но будут доступны для бесплатного пользования всем библиотекам — членам ЦНБ; для сторонних пользователей доступ будет предоставляться на коммерческой основе. Такие коммерческие ресурсы позволяют в ходе программы оцифровки использовать защищённые «копирайтом» материалы, позволяя выкупать права у издателей. Таким образом, выработана оптимальная модель взаимодействия библиотечного и издательского сообщества, которая позволяет части членов ЦНБ (меньше половины общего количества) финансировать программу оцифровки материалов для всех биб-

лиотек, состоящих в ЦНБ. Это особенно необходимо для небольших учреждений и для тех организаций, которые испытывают проблемы с финансированием. Кроме того, в ЦНБ сформирован совещательный комитет по выбору приоритетов в реализации данной программы и формировании новых оцифрованных фондов и коллекций. Этот комитет постоянно взаимодействует с экспертами при выборе конкретных названий для создания персональных коллекций. Как уже было сказано, все новые цифровые файлы поступают на долгосрочное хранение в ЦНБ и используются в цифровых гуманитарных исследованиях.

Работа над «Архивом мировой прессы» совместно с ЦНБ не только трансформировала коммерческую программу оцифровки газет, но и привела к созданию новой модели сотрудничества.

Речь идёт об обеспечении максимально широкого доступа к редким материалам с помощью комбинации политики свободного доступа к ресурсам, уже перешедшим в общественное достояние, с наиболее экономически выгодными условиями платного доступа, определяемыми коллективным участием в нём консорциума библиотек.

Также эта модель обеспечивает библиотекам и учёным активное участие в отборе и подготовке требующихся им ресурсов для долгосрочной сохранности и гуманитарных исследований.

Реализация столь амбициозного проекта требует особых решений в сферах менеджмента, логистики, контроля качества оцифровки, взаимодействия с издателями и обеспечения долгосрочного использования результатов реализации программы оцифровки. По ходу реализации сотрудничества «Ист Вью» со Стэнфордом (уже четвёртый год) и с ЦНБ (завершён первый год совместной работы) накоплен опыт решений в вышеперечисленных сферах совместной деятельности.

Логистика

Основными проблемами оцифровки больших массивов газет считаются следующие: метаданные, выгрузка и обработка, совместное использование конечного ресурса.

Контент фондов Стэнфорда и Гуверовского института собирался и каталогизировался в течение столетия. Менялись правила транслитерации названий, особенно на редких языках, другие ключевые метаданные — к примеру, названия городов. В результате возникало дублирование титульных данных, а в некоторых случаях ключевые метаданные вообще терялись, либо отсутствовали единообразные идентификаторы для выявления дублетных записей.

По мере расширения совместной работы всех партнёров метаданные становились проблемой, так как разные организации нередко использовали различные системы транслитерации и обозначения названий городов при каталогизации фондов. Названия самих изданий также вызывали трудности при их обработке: как для западной газетной культуры, так и для восточной традиции характерно то, что издания имеют похожие (или даже одинаковые) названия. Например, на Западе распространены многочисленные «Дейли Ньюс» или «Таймс», на арабском Востоке — «Акбар», в Китае «Ли Бао» и пр. Поэтому всякий раз возникает вопрос: это действительно разные издания, или просто название продублировано при каталогизации, или это газеты с повторяющимися названиями, выпускаемые одним и тем же издателем в разных городах? Такие вопросы нужно было обязательно решать при определении, кому принадлежат права на данное издание.

Также «Ист Вью» широко привлекала экспертов каталогизации на соответствующих языках для того, чтобы удостовериться в уникальности конкретного издания и затем найти его издателя. В итоге «Ист Вью» затратила шесть

месяцев на эти операции по поиску и очистке данных от дублетных записей, исправлений и обновлений метаданных, а также по регулированию проблем издательских прав и нахождения новых или дополняющих уже используемые фонды и коллекции.

И всё-таки, даже с учётом перечисленных проблем, чистка метаданных куда менее болезненна, чем физическая доставка уязвимых и хрупких печатных материалов для сканирования — иногда с других концов страны. По подсчётам «Ист Вью», объёмы газетного фонда Стэнфорда и Гуверовского института превышали 25 млн страниц текста, причём большая часть этих газет хранилась в виде обёрнутых пачек (общее количество их было не менее 25 тыс.) либо в архивных коробках. Каждая пачка весила около 10–15 фунтов (1 фунт — 454 грамма), а вся обычная библиотечная коллекция — гораздо более 250 тонн. Поэтому трудности были не только транспортные, но и организационные: нужно было тратить немало времени на поиск и инвентаризацию требуемого материала с проверкой возможных его повреждений (а также иной раз с установкой пропаж). Затраты на эту работу и транспортировку оказывают наибольшее воздействие на совокупный бюджет проекта оцифровки. «Ист Вью» предвидела появление этих проблем, а потому затратила немало времени на поиск уже существующих и подходящих для оцифровки копий (в микроформных или электронных форматах) вместо оригиналов. К счастью, миллионы печатных страниц уже существовали в разных фондах в оцифрованном виде — для последующего открытого доступа или для создания коммерческих баз данных. Эти ресурсы были выведены из списка приоритетных для будущей оцифровки. Кроме того, уже существовали миллионы газетных страниц на микрофильмах для архивного хранения, и обычно они достаточно хорошего качества, т. е. могут использоваться в качестве оригиналов для сканирования. ЦНБ обеспечил для «Ист Вью» доступ к максимально широкому фонду таких копий печатных газет на различных носителях. Но всё-таки ещё больший объём газетных фондов требовал сканирования именно печатных оригиналов. Это было необходимо для ликвидации лакун в фондах микрофильмов либо для достижения требуемого качества оцифровки в том случае, если старый микрофильм не годился в качестве оригинала, либо если редкое издание или ресурс никогда прежде не микрофильмировали и не оцифровывали. Надо сказать, что в фондах Стэнфорда или Гуверовского института имеются и такие издания, которые считаются единственными не только в Северной Америке, но и во всём мире.

Логистические проблемы не заканчиваются с доставкой материалов в нужное место. «Ист Вью» оцифровывает все газеты по единым стандартам, базирующимся на тех, которые были разработаны для Национальной программы оцифровки газет (NDNP) Библиотеки Конгресса США. Конечный продукт содержит высококачественные цветные изображения в теговом формате файла изображений (TIFF) с разрешающей способностью 400 dpi (точек на дюйм), файлы XML METS/ALTO (формат кодирования метаданных и их передачи), JPEG2000 (формат сжатия неподвижных цифровых изображений) и PDF (формат представления преимущественно текстовых документов) с функцией поиска. Их суммарный объём составляет многие петабайты (т. е. миллионы гигабайтов) данных. Наряду с обеспечением долгосрочной сохранности всех этих файлов и предоставления к ним доступа для научных исследований «Ист Вью» предоставляет копии этих файлов всем участникам проекта, включая Стэнфорд и ЦНБ.

Компания тесно взаимодействует с разными экспертами сканирования, цифровой конверсии и создания различных платформ. Фактически работа даже с единственной газетой ведётся в разных уголках мира. Она может начинаться в Калифорнии (Стэнфорд) или в Иллинойсе (Чикаго), как и в любом институте-партнёре в США, но вполне возможен её старт в Японии или Китае. А затем эта работа продолжается уже в Миннеаполисе при поддержке в Москве и Киеве, в Нью-Брунсвике (штат Нью-Джерси) и Пномпене, а то и в новозеландском Веллингтоне. Конечные полученные файлы отправляются затем в Миннеаполис, Стэнфорд и Чикаго.

При полномасштабном режиме работы объём оцифровки составляет 500 тыс. страниц ежеквартально. Скорость электронной доставки, т. е. пересылки и выгрузки этих файлов, сильно варьируется, но в целом «не поспевает» за такими объёмами. Поэтому «Ист Вью» при поддержке внешних партнёров типа DDD, CCS docWorks (специалистов по конверсии и управлению контентом), «Amazon Web Services» и других решили задействовать систему хранения, основанную полностью на «облачных» технологиях. Отсканированные «Ист Вью» в Миннеаполисе, Нью-Брунсвике, Киеве или Москве материалы, как и полученные от других партнёров или издателей со всего мира, загружаются в виртуальные «ящики» системы «Amazon S3», которые автоматически передаются в DocWorks-станции, управляемые техническим персоналом в Пномпене. Эксперты «Ист Вью» получают доступ к этим «ящикам» в режиме QC («быстрое соединение») для проверки качества материалов. После проверки эти файлы выгружаются на платформы в Веллингтоне и отправляются в другие «ящики», доступ к которым получают специалисты Стэнфорда и ЦНБ. Они производят уточнение и «чистку» метаданных, в том числе информации по депозитариям конкретных материалов, для гарантии того, что каждый партнёр получит нужный ему контент.

Процесс оцифровки и оптического распознавания символов (OCR)

Весной 2017 г. «Ист Вью» связалась с DDD по поводу возможного его участия в вышеописанном проекте. После пробного ввода и вывода данных с жёстких дисков стало очевидно, что такая технология не рассчитана на требуемые масштабы, поэтому было решено попробовать «облачное» хранение.

DDD, являясь пользователем программного обеспечения системы CCS docWorks с 2007 г., применила эту технологию наряду с собственными наработками при их пробном испытании и сравнении. «Ист Вью» имеет также опыт давнего сотрудничества на уровне экспертизы с веб-службами компании «Амазон» — AWS. CCS уже много лет поставляет централизованное программное обеспечение для оцифровки печатных материалов различным пользователям-библиотекам и имеет по всему миру свои сервисные бюро с начала 2000-х гг. Однако это обеспечение не связано с «облачными» технологиями, поэтому участники данного проекта решили рискнуть и сосредоточиться исключительно на сотрудничестве с AWS.

Риск был связан с масштабами оцифровки: такие объёмы никогда прежде не поступали в «облачное» хранение. Вводимые изображения иногда достигали объёма 400 мегабайт, а производные выводные данные могли превышать исходники по размеру в 2,5 раза и более. Перемещая такие объёмы информации для ввода, размещения и вывода из «облака», мы были здесь пионерами, особенно при взаимодействии с партнёрами, разбросанными по всему миру. К счастью, перевод программного обеспечения CCS на использование «облачных»

технологий AWS оказался довольно лёгким и практически безболезненным. После короткого тестового периода удалось запустить систему на полную мощность, и она оказалась эффективной и быстродействующей.

Как CCS, так и AWS предлагают пользователям модель оплаты их услуг по реальному объёму использования, что значительно уменьшает клиентам размер первоначальных капиталовложений и непроизводительные расходы. Главная составляющая цены приходится на стоимость хранения файла в AWS. Объём и сроки требуемого хранения определяются временем между загрузкой и приёмкой обработанных ресурсов.

Фонды Стэнфорда и Гуверовского института содержат издания более чем из 125 государств. Фонды ЦНБ и его членов также включают издания из многих стран. Поэтому мы имеем дело с огромным разнообразием языков и шрифтов — от латиницы с использованием гарнитур типа антиква и готического шрифта до сложнейших китайских (традиционный и упрощённый) и японских[3] комбинаций; здесь же корейский, русский, испанский, арабский, урду, пушту.

Имеется также множество оригиналов с невысоким качеством изображения текста по причине повреждения бумажных носителей или плохого качества микрофильмов, если они использовались для оцифровки вместо исходных печатных изданий. В других случаях, даже имея дело с высококачественным оригиналом для сканирования OCR, получаемые цифровые ресурсы оказывались непригодными, так как программное обеспечение для некоторых языков не было столь точным, как для западных, пользующихся латинской графикой.

Форматы печатных газет также создавали в работе немало трудностей. Форматы газетных полос варьируются очень широко и даже для одной и той же газеты нередко меняются. Есть и другие проблемы — к примеру, с вёрсткой текстов. Японские газеты обычно читаются сверху вниз, но заголовки статей часто пишутся слева направо; таков же бывает и порядок размещения иллюстраций. При этом заголовки статей не всегда помещают над текстом, а верстают в другом месте.

Особенно трудно обрабатывать арабские тексты, поскольку буквы там соединены друг с другом, а некоторые имеют, к тому же, несколько вариантов начертания в зависимости от местоположения в слове. Арабский язык содержит ряд диакритических знаков, которые в современных публикациях, например газетных, нередко опускаются, но обязательно присутствуют в текстах на классическом арабском языке и особенно в исторических документах. Кроме того, существует множество диалектов, в каждом есть особенные слова, не используемые в других регионах. Некоторые языки могут быть представлены в арабской шрифтовой графике, но с добавлением букв из алфавитов других языков, например урду, пушту, персидского, турецкого, оттоманского или пенджаби. Для ряда из них не существует программного обеспечения для сканирования с распознаванием (OCR).

В подобных случаях возможно обучение персонала и переналадка программного обеспечения, в частности с помощью имеющейся в свободном доступе программы «Tesseract». Но такая работа требует многих затрат, в том

[3] Современная японская письменность включает в себя две традиционные слоговые азбуки (катакана и хирагана) и систему кандзи, основанную на иероглифах китайского происхождения, вместе с которыми используется фуригана (система фонетических подсказок), а также ромадзи (элементы латинской графики), римские и арабские цифры, греческие буквенные обозначения и пр. Все эти системы могут использоваться в пределах не только одной газетной полосы, но даже одной заметки, а то и заголовка. — *Примеч. переводчика.*

числе денежных. Проект сбора и архивирования машиночитаемых научных текстов «Открытая инициатива исламизации текстов» (OpenITI) охватывает лишь некоторые языки с арабской графикой. По поводу тамильского языка многие культурно-исторические институции в течение ряда лет высказывали настойчивые пожелания относительно разработки программы с OCR, но надёжной работающей программы так и не появилось.

Мы не затрагиваем здесь проблемы низкого качества многих изображений (будь они в самих печатных газетах или в их микрофильмах) потому, что они относятся не только к изданиям на арабском или другом языке со сложной графикой, но и к любому газетному оригиналу для сканирования, с любым алфавитом и начертанием букв. Члены и участники проекта IMPACT (компании DDD), ныне именуемого «Центром компетенций IMPACT», создали ряд контролирующих механизмов, тестовых наборов, эталонных критериев и словарей для языков с латинским алфавитом.

Издатели и «копирайт»

Как уже указывалось, оцифровываемые фонды в основном относятся к изданиям XX в. приблизительно из 125 стран. При этом большая часть контента остаётся под защитой «копирайта». Хотя у конкретного учреждения-депозитария, где хранится то или иное издание, обычно имеются права на его оцифровку для последующего сохранения, эти права не соответствуют главной задаче описываемой программы — обеспечению максимально широкого доступа к материалам для научных исследований. «Ист Вью» должна в любом случае уважать законные права владельцев «копирайта». Более того, библиотеки, соблюдающие принципы авторского права, вправе ожидать, что поставщики агрегированных баз данных полностью обеспечат защиту оцифрованных материалов, охраняемых «копирайтом». Именно такая функция была ключевой, когда провайдер в лице «Ист Вью» играл роль посредника между издателями и учёным сообществом.

Поэтому «Ист Вью» организовал изучение международного законодательства в сфере охраны интеллектуальной собственности как в общемировом масштабе, так и в отдельных странах. В разное время и в разных местах действовали и действуют неодинаковые законодательные нормы, и все они должны соблюдаться с учётом того, где и когда состоялась публикация и того, где располагается распространитель контента, и того, куда поставляются материалы.

Учитывая объёмы выпуска и распространения газетных публикаций по всему миру, проблемы «копирайта» были самыми сложными для реализации данной программы. Хотя большинство стран и приняли международную Бернскую конвенцию, очень многие дополнительно вводили национальные правила и ограничения, которые шли намного дальше границ исходного соглашения об авторских и издательских правах 1886 г. Кроме того, от времени подписания данной конвенции конкретной страной зависела ситуация с защитой такого «копирайта» до и/или после присоединения этой страны к Бернской конвенции. К тому же, количество международных договоров и конвенций в этой сфере непрерывно возрастает. В качестве примера укажем на договоры стран Европейского Сообщества либо США — Мексики — Канады. Иными словами, проблемы «копирайта» постоянны, да ещё и быстро трансформируются.

В итоге каждое газетное название — более 2 тыс. из Стэнфорда и Гуверовского института плюс более 10 тыс. из ЦНБ — должно быть тщательно и отдельно проверено на предмет действующих ограничений в его использовании.

Часть таких изданий выдвигается на первый план, поскольку срок защиты «копирайта» для них истёк, либо они прекратили издаваться сто лет назад без какой-либо преемственности, а потому могут быть обработаны как части фондов открытого доступа. Другую группу составляют издания, которые продолжают выходить и их легко найти: это позволяет «Ист Вью» выкупать права у издателя как роялти от коммерческого распространения.

Однако между двумя названными группами газетных изданий есть промежуточная группа «сирот» — это названия, которые прекратили выпуск, никем не продолжаемые и не имеющие ныне здравствующих известных владельцев. «Сиротами» обычно оказываются редкие или/и малотиражные издания, выступавшие против правящего порядка или окончившие существование на стороне побеждённых в гражданской войне. А как раз такие материалы могут быть важны и чрезвычайно интересны для научных исследований и потому требуют глубокого анализа применительно к такой проблеме, как баланс между риском нарушения прав на интеллектуальную собственность и ценностью информации для науки. Здесь «Ист Вью» принимает решения исключительно на индивидуальной основе с максимальным уважением ко всем обнаруженным правообладателям. Если возникает необходимость, компания платит издателям за оцифровку даже прекратившихся изданий, например «Новое русское слово» или «Москоу Ньюс».

Финансирование и доступ

Как и логистика с OCR и копирайтные затраты, программы оцифровки газетных массивов везде требуют больших денежных расходов. Это касается даже коллекций с открытым доступом. Поэтому многие создатели баз данных фокусируются на наиболее тиражных названиях газет текущего выпуска, что сулит наибольшие прибыли в условиях платного доступа, либо, наоборот, сосредоточивают внимание на материалах, уже не защищённых «копирайтом», т. е. не требующих выплат их издателям, что тоже позволяет соблюдать свою выгоду. Но оба подхода снижают научную ценность представленного контента, а также приводят к явному доминированию западного взгляда на исторические и другие события при практически полном игнорировании мнений другой половины земного шара. Отсюда огромные информационные пробелы в материалах по исследованию XX в.

«Ист Вью» стремится отрегулировать такой дисбаланс с помощью крупных капиталовложений в процессы оцифровки возможно большего количества разных газет и обеспечения научному сообществу максимально широкого доступа к подобным ресурсам.

Благодаря сотрудничеству с ЦНБ «Ист Вью» удалось ввести в научный оборот множество «сиротских» газетных изданий со всего мира. Члены ЦНБ согласились профинансировать оцифровку 3 млн газетных страниц из 6 коллекций с открытым доступом. Как участник этой программы ЦНБ, «Ист Вью» также оцифровал 1,5 млн страниц газетного контента, ещё защищённого «копирайтом». Этот ресурс был затем передан всем членам ЦНБ, независимо от их участия или неучастия в финансировании совместной программы «Ист Вью» и ЦНБ.

Отдельные издатели также получат выгоду от преобразований газетного контента, так как ожидают получения роялти за платный доступ, когда он будет введён хотя бы для некоторых газетных изданий. Издатели при этом справедливо рассчитывают на то, что такой процесс может быть долговременным. В то же время они нередко колеблются, принимая решение войти (или не войти) в число участников больших сводных коллекций, поскольку это сулит потен-

циальную конкуренцию с другими издателями. Ситуация требует, чтобы многие названия выступали в качестве единственных в своём роде коммерческих продуктов. По факту именно издатель определяет модель распределения своего контента среди пользователей. Никто из издателей не разрешит свободный доступ к своим материалам по всему миру, немногие дадут согласие войти в сводные коллекции для обеспечения доступа к ним на базе консорциумов пользователей, но большинство издателей будут ожидать, что к их материалам будут относиться как к уникальным и единственным. Это должно выразиться и в статусе такого контента, и в особенностях условий его распространения.

Данная статья доказывает необходимость прочного сотрудничества коммерческого и научного секторов для поддержки и реализации программ оцифровки массивов газетных изданий. Коммерческие вендоры типа «Ист Вью» незаменимы при наведении мостов между научными структурами и издателями. Другие вендоры типа DDD и CCS играют большую роль в экспертной оценке контента и логистике. Академические и научные учреждения — депозитарии фондов также вносят огромный вклад в финансирование оцифровки, будь то на основе отдельных заявок конкретных учреждений либо долевое финансирование на базе потребителей консорциумов, либо финансирование для передачи контента в открытый доступ. В большинстве случаев отдельные научные учреждения не имеют нужного финансового и логистического опыта для реализации масштабных программ такого рода своими силами. Также у них будут проблемы с партнёрством издателей, и в результате либо доступ к контенту «замкнётся» внутри стен только этой организации, либо оцифровывать можно будет только контент, не защищённый «копирайтом». Опыт доказывает, что успешная реализация таких грандиозных программ, как «Архив мировой прессы», возможна только на базе широкого партнёрства отдельных научно-информационных учреждений и их объединений (например, в виде библиотечных консорциумов) и коммерческих вендоров наподобие «Ист Вью». Будущее этой программы видится в расширении количества её участников и обещает в ближайшие десятилетия трансформировать многие научные исследования.

Библиографический список

1. Stanford Libraries https://library.stanford.edu/.

2. Hoover Institution Library & Archives https://www.hoover.org/library-archives.

3. Center for Research Libraries http://www.crl.edu/.

4. East View Global Press Archive https://www.eastview.com/resources/gpa/.

5. Digital Divide Data https://www.digitaldividedata.com/.

6. CCS Content Conversion www.content-conversion.com/.

7. IMPACT Centre of Competence https://www.digitisation.eu/.

8. Open Islamicate Texts Initiative OpenITI https://iti-corpus.github.io/).

Перевёл с сокращениями

Константин Михайлович Сухоруков

Российская книжная палата (филиал ИТАР–ТАСС), заместитель директора, кандидат исторических наук, Россия, Москва, e-mail: a-bibliograf@mail.ru

Konstantin Mikhailovich Sukhorukov

Russian Book Chamber (branch of ITAR–TASS), deputy director, candidate of historical sciences, Russia, Moscow, e-mail: a-bibliograf@mail.ru