УДК 003 Вестник СПбГУ. Сер. 13. 2014. Вып. 4

О. И. Редькин, О. А. Берникова

ПРОБЛЕМЫ ОЦИФРОВКИ И КАТАЛОГИЗАЦИИ АРАБОГРАФИЧЕСКИХ РУКОПИСЕЙ

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7/9

В статье представлен анализ различных методов каталогизации арабографических рукописей. Рассмотрев традиционные подходы к описанию манускриптов, а также существующие сегодня электронные коллекции арабских рукописей, авторы предлагают свой подход к дигитализации рукописного материала. Данный подход основан на разработке так называемого цифрового паспорта рукописей, который позволит осуществить их классификацию, сделать выводы относительно авторства, определить рукописную школу, а также установить датировку. Библиогр. 17 назв.

Ключевые слова: оцифровка рукописи, каталогизация, арабографический манускрипт, кодикология.

PROBLEMS OF DIGITIZING AND CATALOGING ARABIC MANUSCRIPTS

O. I. Redkin, O. A. Bernikova

St. Petersburg State University, 7/9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

The article presents a thorough analysis of different approaches to cataloging manuscripts, based on the arabic script. After reviewing the traditional methods of manuscript description and analyzing available digital collections of arabic manuscripts, the authors proposed their own technology of digitizing handwritten materials. This approach relies on the development of the so-called digital "passport" of manuscripts that in its turn will help to simplify cataloging process in order to define the authorship, to attribute it to a certain handwriting school and to verify the dating of a manuscript. Refs 17.

Keywords: digitization of manuscripts, manuscript cataloging, arabic script, codicology.

Введение

Средневековые арабографические рукописи представляют собой один из важнейших элементов арабо-мусульманского наследия, источник по истории Ближнего Востока и Северной Африки, а также Средней Азии, Кавказа, Европы и России, свидетельством чему являются работы В. В. Бартольда, В. И. Беляева, А. Б. Халидова, И. Ю. Крачковского, считавшего «непрерывность научного преподавания арабского языка в высшей школе и постепенное расширение доступных для исследования рукописных материалов» необходимыми условиями для развития арабистики [1, с. 70–71]. При этом значимость изучения арабских манускриптов не ограничивается лишь историографией и востоковедением, они также могут быть источником информации и для специалистов в области других наук — медицины, математики, философии, астрономии, истории естествознания.

Целью данной статьи является описание авторской технологии автоматизированной каталогизации арабографических рукописей на основе определения ключевых распознавательных маркеров. Помимо этого рассматривается принципиальная возможность проведения исследования междисциплинарного характера при работе с рукописным материалом, сочетания методов кодикологического и компьютерно-

го анализа метаданных (методов оптимизации и визуализации). Рассматриваются наиболее распространенные технологии формирования электронных рукописных каталогов.

Каталогизация рукописей: традиции vs технологические решения

Формирование рукописных фондов в Санкт-Петербурге восходит к XVIII в., однако их систематическое изучение началось лишь с приездом в столицу империи академика Х. Д. Френа, созданием здесь Азиатского музея, а также приобретением новых коллекций манускриптов [1].

Сегодня, как и прежде, изучение и описание рукописного арабографического наследия является одной из актуальных задач арабистики. При этом следует отметить, что кодикологические методы, используемые при исследовании рукописей на европейских [2, с. 23–29] и некоторых восточных языках, применяются еще не в полной мере при изучении их арабских аналогов.

Несмотря на то что традиция изучения арабских рукописей в нашей стране [3] и в Европе в целом имеет давнюю историю, методы описания и исследования манускриптов практически не изменились.

Даже предварительное знакомство с работами, посвященными изучению арабского рукописного наследия, показывает, что принципы формального, а также текстологического и, в более узком смысле этого слова, лингвистического анализа рукописного текста варьируются незначительно и основаны на четко выверенной и проверенной временем парадигме описания манускриптов. Такая парадигма предусматривает их хронологизацию, лексический и терминологический анализ, соотнесение с одной из рукописных школ, изучение исторического или лингвистического контекста, сопутствующего их созданию или переписке.

Как правило, в исследованиях такого рода рассматриваются особенности языка рукописного памятника, почерка, свойства чернил, тип переплета, тематика, а также элементы паратекста — колофоны (сведения об авторе, месте и времени переписки, имени переписчика/заказчика в конце рукописи) и субскрипции (то же в начале рукописи), инципиты и эксплициты (формулы начала и конца текста), наличие печатей. При этом анализ арабских манускриптов основывается в значительной степени на субъективных оценках исследователя, которые, в свою очередь, зависят от его квалификации, опыта и знаний.

Что касается метрических данных, то, как правило, приводятся лишь сведения о количестве и размерах страниц, параметрах переплета, количестве строк на странице. Более детальные характеристики, связанные с так называемой ритмикой рукописного текста, за редкими исключениями остаются вне поля зрения исследователей.

Вместе с тем такие показатели, как цветовые характеристики текста, его «ритмика» [4], индивидуальные особенности почерка переписчиков или авторов манускриптов в случае, если исследователю посчастливится работать с автографом, а также ряд других могут быть объектом описания с использованием формальных цифровых параметров.

За последние два века сложилась и определенная традиция каталогизации рукописных фондов, осуществляемая в зависимости от контента и тематики, либо авторства сочинений. Что касается публикаций отдельных рукописей, то в них, как правило, представлены текст сочинения на языке оригинала, перевод, а также лингво-исторический или филологический комментарии. Кроме того, публикации могут быть снабжены изображениями отдельных фрагментов оригинальных текстов рукописей. Рукописные каталоги до недавнего времени были представлены на бумажных носителях, и лишь в последние десятилетия все большее распространение получили их электронные аналоги, размещаемые на сайтах библиотек или научно-исследовательских центров, которые выгодно отличаются по ряду показателей от изданных в типографиях аналогов.

Проблемы терминологии

Использование информационно-коммуникационных технологий при создании электронных рукописных каталогов влечет за собой появление новой терминологии для описания данного процесса. В последнее время широкое распространение получил термин «дигитализация», или «оцифровка» рукописей, который трактуется как цифровая компьютерная обработка манускриптов. Дигитализация арабского рукописного наследия может внести существенные коррективы в процесс описания рукописей и помочь в разработке электронных он-лайн каталогов и, в известной степени, оптическом распознавании символов.

Еще сравнительно недавно под оцифровкой документа понималось создание его альтернативной, предназначенной для хранения копии [5]. В настоящее время существует несколько различных толкований данного термина.

В упрощенном виде под «оцифровкой» нередко понимается создание электронной копии изображения объекта, что при работе с рукописными документами означает их сканирование и запись полученных данных в форматах BMP, JPG, JPEG или других типах файлов на электронных носителях или их размещение на сайтах академических учебных заведений. Такого рода альтернативные коллекции рукописей составляют сегодня абсолютное большинство среди собраний манускриптов, имеющихся в виртуальном пространстве.

Значительно реже под оцифровкой понимается не только создание электронной копии документа, но и последующая его сегментация на отдельные символы (буквы). В этом смысле термин «оцифровка» применительно к существующим электронным коллекциям манускриптов используется крайне редко.

Следует отметить, что, несмотря на то что в последние годы было разработано значительное количество предназначенных для этой цели программных продуктов, в том числе и достаточно успешных в коммерческом плане, проблема оптического распознавания текста применительно к арабскому материалу не утратила своей актуальности и по сей день. Специальное тестирование приложений, созданных ведущими разработчиками лингвистических программных продуктов, также свидетельствует о том, что в настоящее время отсутствуют предназначенные для оптического распознавания печатного текста программы, которые отличались бы достаточно высоким уровнем эффективности и функциональности. Причины этого были

подробно рассмотрены авторами настоящей статьи в опубликованных ранее работах [6; 7].

Создание эффективной программы для распознавания арабского печатного, не говоря уже о рукописном, текста представляет собой более сложную проблему, нежели разработка аналогичного продукта, предназначенного для работы с текстами на основе кириллицы или латиницы.

При оптическом распознавании следует учитывать и то, что тип письменности рукописных документов варьируется в зависимости от традиций определенной рукописной школы, а также может включать в себя индивидуальные особенности почерка авторов. Проблема распознавания становится еще более сложной, когда приходится иметь дело с дополнительными «шумами», такими как комментарии переписчиков, дефекты письма, повреждения используемого материала, а также лакуны и пропуски, равно как и более поздние по времени дополнения к первоначальному тексту. Все это делает адекватную идентификацию арабских письменных текстов крайне затруднительной.

Перевод арабского рукописного текста в цифровой формат с возможностью его последующей обработки и осуществлением функции поиска в массиве контента делается преимущественно в ручном режиме, и, в силу высокой трудоемкости, такой метод оцифровки распространен весьма мало. В данном контексте термин «оцифровка» по отношению к существующим электронным коллекциям манускриптов имеет ограниченное количество примеров.

Ниже будут рассмотрены некоторые базы данных арабографических рукописей и используемые в них технологические решения, позволяющие осуществлять поиск отдельных слов в контенте рукописного материала.

Нередко термин «оцифровка» используется и при описании методов, применяемых в работе с историческими документами. В этом случае речь идет о создании электронной копии рукописи, а также ее аннотации и описании в автоматизированном режиме с целью более эффективной и быстрой каталогизации. Все это требует разработки сложной информационной системы, координирующей связь между изображением документа и комплексом имеющихся и вновь получаемых метаданных, что близко к методике кодикологического исследования.

Понятие «метаданные» в контексте изучения рукописного наследия также весьма вариативно. Так, «классическое» толкование этого термина подразумевает определение традиционных характеристик рукописи (имя автора, материал, почерк и т. д.) и проведение на их основе дальнейшей каталогизации. Данная трактовка не включает в себя осуществление дигитализации исходного материала.

В настоящее время термин «метаданные» трактуется гораздо шире и понимается как возможность соотнесения и сопоставления основных параметров рукописей при создании электронных коллекций манускриптов.

Иными словами, метаданные — это «информация» об «информации». Именно благодаря метаданным осуществляется автоматизированная фильтрация и поиск данных. Метаданные — это «структурированная информация, которая описывает, объясняет, находит или упрощает извлечение, использование или управление информационным ресурсом» [8, p. 2]. В ходе создания той или иной цифровой рукописной коллекции определяется перечень конкретных характеристик, которые будут использованы в качестве маркеров во время автоматизированного поиска.

В настоящее время в мировой практике сложились единые принципы использования тех иных метаданных при составлении электронных рукописных коллекций [9], что позволяет обеспечить более широкий доступ к историческим документам.

Электронные коллекции арабографических рукописей: современные решения

На протяжении последних полутора десятилетий активно разрабатывались и применялись на практике технологии оцифровки и каталогизации арабографических рукописей. В результате было создано значительное количество рукописных коллекций, как размещенных в открытом доступе, так и требующих авторизации пользователей. Например, на веб-странице библиотеки Мичиганского университета (г. Анн Арбор, США) можно найти ссылки на сотни электронных коллекций арабомусульманских манускриптов в различных регионах мира [16].

Технология составления представленных коллекций, равно как и возможности автоматизированного поиска необходимой информации, варьируются. В качестве примера наиболее распространенной модели формирования и хранения манускриптов можно привести рукописную коллекцию, созданную в Университете Ан-Наджах (г. Наблус, Палестина) [11], представляющую собой сканированные полнотекстовые копии рукописей и их электронный каталог. При этом система позволяет осуществлять автоматизированный поиск по одному из следующих параметров: название рукописи, ее автор, дата переписки, номер в каталоге, тематика, время создания (относительно периода жизни автора), имя переписчика. Все эти данные введены в систему в ручном режиме. Несмотря на то что в коллекции представлены лишь электронные копии изображений (что не позволяет осуществлять поиск отдельных слов в тексте рукописи), работать с такого рода цифровой коллекцией достаточно удобно: имеется возможность распечатать фрагмент рукописи в формате PDF с автоматическим указанием на первоисточник. Хорошо продуманы также дизайн и навигация по сайту.

Как правило, проекты создания крупных рукописных коллекций являются примерами совместной международной деятельности библиотек, институтов, фондов. Так, сотрудничество Библиотеки Принстонского университета (США), Свободного университета Берлина (ФРГ) и культурного фонда Имам Зайд ибн Али (IZbACF) в Сане (Йемен) [12] позволило создать электронную коллекцию, которая является частью Принстонской цифровой библиотеки исламских рукописей [13].

Другим примером успешной реализации проекта по дигитализации рукописного наследия стала разработка, представленная так называемым Консорциумом по каталогизации арабских рукописей, который включает в себя библиотеку Wellcome в Лондоне (Великобритания), Александрийскую библиотеку (Египет) и Департамент цифровых гуманитарных наук (Королевский колледж, г. Лондон, Великобритания) [15]. Навигация по веб-сайту данной коллекции достаточно удобна для пользователей. Как рукопись, так и ее метаданные могут быть сопоставлены и проанализированы одновременно. Разработчики создали унифицированную поисковую систему, позволяющую обрабатывать текстовую информацию, содержащуюся в следующих фрагментах рукописи: инципите, колофоне, основном тексте, комментариях и информации о происхождении рукописи. Несмотря на уникальный междисциплинар-

ный подход, который был применен при разработке данной коллекции, необходимо отметить следующее:

1. Бóльшая часть представленных в коллекции рукописей сходна по тематике и представляет собой сочинения по медицине, что значительно упрощает автоматизированный поиск.

2. В аннотации к коллекции отмечается, что применяемые технологии позволяют осуществлять полнотекстовый поиск. Если такое утверждение верно, то, учитывая значительный объем рукописной коллекции, оно свидетельствует о решении сложной задачи распознавания рукописного текста, что весьма сомнительно. Известно, что программные приложения по распознаванию арабографических текстов работают несовершенно, даже в контексте обработки печатного материала. Распознавание рукописных документов, которые включают в себя индивидуальные особенности почерка авторов, — еще более сложная задача, не говоря уже о дополнительных «шумах», как то комментарии переписчиков, дефекты письменных материалов, а также лакуны и пропуски, дополнения к первоначальному тексту. Все эти особенности делают правильную идентификацию арабских письменных текстов крайне затруднительной. Успешным опытом создания технологии автоматизированного поиска в рукописном тексте является разработка рукописной базы данных топонимов (названий арабских городов / деревень) [15]. Можно предположить, что успешность такого рода исследования обеспечена спецификой и определенным количеством лексического контента.

С этой точки зрения технология поиска, разработанная для коллекции «Арабские рукописи онлайн» библиотеки Wellcom, вызывает ряд вопросов, в частности: действительно ли она предполагает поиск по содержимому текста всей рукописи (полнотекстовый поиск) или применима лишь к фрагментам рукописей, текст которых введен в компьютер в ручном режиме?

Приведенные выше примеры электронных коллекций арабографических манускриптов по сути являются попытками создания электронных баз данных с возможностью автоматизированного поиска. При этом значительная часть самой работы по дигитализации и каталогизации имеющихся манускриптов проводится вручную. Предлагаемые авторами настоящей статьи методы оцифровки и классификации рукописей значительно отличаются от технологий, используемых в WAMCP, и направлены на разработку технологии автоматизации процесса каталогизации сканированных рукописей с использованием метаданных.

Анализ имеющихся в этой области разработок позволяет выделить лишь эксперимент группы исследователей из Каирского университета, разрабатывавших технологию, направленную на решение аналогичных задач. Основной целью их проекта являлась классификация исторических документов относительно трех исторических периодов: современного, османского и мамлюкского [16].

Инновационные методы каталогизации

До недавнего времени под цифровой обработкой рукописей подразумевалось их сканирование и дальнейшее форматирование полученного изображения, а также, в случае необходимости, размещение его в сети Интернет или создание локальных баз данных. При этом изображение, как правило, было представлено в формате

2D. Вместе с тем необходимость ввода в научный оборот дополнительного материала требует и новых технологических решений, которые не ограничиваются предварительным исследованием и обработкой документов на основе арабской графики и артефактов. Все это возможно лишь при условии совместной работы специалистов в области компьютерного программирования и арабистов — исламоведов, филологов и лингвистов. Одним из решений такого рода должно стать создание автоматизированной системы распознавания арабского текста, в том числе и рукописного.

Эффективность существующих продуктов распознавания для арабского языка, таких как Sakhr [17], в значительной степени зависит от структуры анализируемых материалов, они функционируют с минимальным количеством ошибок лишь при условии работы с «идеальным» текстом (набранным с помощью одного из наиболее распространенных шрифтов, лишенным огласовок и т. д.).

Очевидно, что проблема распознавания символов в тексте на арабском языке сложнее, чем в текстах на латинице или кириллице. Данное обстоятельство во многом обусловлено проблемами как лингвистического, так и технологического характера. Трудность распознавания арабской графики обусловлена и большим количеством дериватов, «слитным» характером письма, допускающим различную длину соединительных линий, возможность реализации точек в стороне от буквы, наличием лигатур, слитным написанием ряда предлогов и частиц.

Электронный «паспорт» рукописи

Актуальная необходимость использования методов компьютерного анализа рукописей обусловлена и такими особенностями рукописного текста, как наличие в нем филиграней, верже и понтюзо, различных типов используемой бумаги. Так, при цифровой обработке можно определить точное количество строк в тексте, под каким углом текст помещается между линиями сетки (верже / понтюзо) на конкретной странице, а также во всей рукописи в целом.

Помимо этого, имеется возможность соотнесения рукописи с конкретной рукописной школой, что на практике возможно при наличии достаточного количества необходимой информации о рукописях с аналогичным типом письма и относящихся к определенной школе переписчиках рукописей, использовавших похожие типы письма. Традиционная классификация почерков при этом недостаточна, поскольку нередко встречаются и «гибридные» варианты письма (например, включающие начертание ف (fā’) / ق (qāf), характерные для *магриби*, при этом сходное с почерком *насх* написание выносных элементов букв).

Только при компьютерном анализе возможна масштабная и объективная систематизация представлений о вариантах написания нижних и верхних выносных элементов, закрытых (م mīm) и открытых (ب bā’) букв, внутрибуквенных просветов, диакритик, огласовок, лигатур. Такого рода характеристики невозможно определить «на глаз» для текста, превышающего один-два абзаца и имеющего сколько-нибудь серьезный объем. Без использования современных технических методов трудно описать и такие особенности арабских рукописей, как аббревиатуры, пояснительные знаки, а также способы их применения и использования в разных регионах исламского мира.

Компьютерная обработка и анализ текста может позволить осуществить автоматизированное выявления таких характеристик, как:

— тип почерка;

— пропорциональное соотношение размеров полей и текста, промежутков между словами и длиной слов;

— соотношение между высотой и шириной букв, расположением диакритик;

— степень наклона почерка;

— сила нажима и, следовательно, тип инструмента (по ширине линий букв);

— наличие различных типов почерков в одном манускрипте;

— процентное соотношение сохранившегося текста и лакун;

— особенности цветовой палитры, иллюстраций, инвентарных пометок и печатей, надписей на полях, цветовых характеристик чернил;

— классификация в зависимости от характера букв, плотности письма и т. п.

Компьютерный анализ позволит также выйти за пределы привычного приблизительного определения почерков «на глаз». Обычная для каталогов характеристика почерков остается в своей основе традиционной (насх, куфи, магриби, насталик и т. п.), достаточно условной («количество точек вдоль линий») и далеко не всегда отражает реальность письма.

Заключение

Приведенный выше комплекс цифровых показателей можно сравнить с данными спектрального анализа, так как он является уникальным для каждого манускрипта. Цифровой анализ рукописей позволит осуществить классификацию вариантов письма на основе объективных (поддающихся цифровому выражению) характеристик. Итогом должен стать спектр вариантов письма, которые могут быть сведены в несколько групп.

Создание такого рода выраженного в цифровых параметрах компьютерного «паспорта» рукописи позволит осуществить ее компаративный анализ в сопоставлении с другими рукописями, осуществить их классификацию, сделать выводы относительно ее авторства, определить, является ли она автографом или списком, наиболее вероятную возможность принадлежности к той или иной рукописной школе, а также помочь в ее датировке. Цифровой «паспорт» рукописей облегчит процесс каталогизации, подготовки их к дальнейшей публикации и использовании в научной работе.

Сказанное выше в значительно степени справедливо и в отношении рукописей, написанных на других языках, графика которых основана на арабском алфавите, например, рукописных текстов на персидском, урду, пушту, дари, кашмири, пенджаби, синдхи, хауса, фула, курдском (в Иране и Сирии), уйгурском, а также ряде других языков, распространенных в Северной и Западной Африке, на Ближнем Востока, южной и юго-восточной Азии.

Литература

1. *Крачковский И. Ю.* Очерки по истории русской арабистики // Избранные сочинения. М.; Л.: Изд-во АН ССС, 1958. Т. 5. 526 с.

2. *Столярова Л. В., Каштанов С. М.* Книга в Древней Руси (XI–XVI вв.) М.: Русский фонд содействия образованию и науке, 2009. 480 с.

3. *Крачковский И. Ю.* Над арабскими рукописями. М.; Л.: Изд-во АН СССР, 1946. 168 с.

4. *Полосин В. В.* Рукописи каллиграфической школы Ибн Муклы (проблема идентификации) // Письменные памятники Востока. СПб., 2004. № 1. С. 160–176.

5. *Zdeněk U.* Digitization is not only making images: manuscript studies and digital processing of manuscripts. Knygotyra, 2008. P. 148–160.

6. *Redkin O. I., Bernikova O. A.* Problems of the Arabic OCR: New Attitudes // Proceedings of the 2013 International Conference on Artificial Intelligence. Las Vegas, USA, 2013. P. 777–782.

7. *Redkin O. I., Bernikova O. A.* On the Optical Character Recognition and Machine Translation Technology in Arabic // Proceedings of the 2012 International Conference on Artificial Intelligence. Las Vegas, USA, 2011. P. 861–867.

8. Understanding Metadata. NISO, 2004. P. 1–20.

9. *Banach M., Shelburne B., Shepherd K., Rubenstein A.* Guidelines for Digitization, Digital Creation and Preservation Working Group. 2010–2011.

10. http://guides.lib.umich.edu/islamicmsstudies/onlinecollections (дата обращения: 08.07.2014).

11. http://manuscripts.najah.edu/ (дата обращения: 08.07.2014).

12. http://wamcp.bibalex.org/about-us (дата обращения: 08.07.2014).

13. http://pudl.princeton.edu/objects/9s1616928 (дата обращения: 08.07.2014).

14. http://pudl.princeton.edu/collections/pudl0032 (дата обращения: 08.07.2014).

15. *Pechwitz M., Snoussi Maddouri S., Märgner V., Ellouze N., Amiri H.* IFN/ENIT-Database of Handwritten Arabic Words // 7th Colloque International Francophone sur l'Ecrit et le Document, CIFED. Oct. 21–23, 2002, Hammamet, Tunis. P. 1–8.

16. *Ahmad M. Abd Al-Aziz, Gheith M., Ayman F. Sayed* Recognition for old Arabic manuscripts using spatial gray level dependence (SGLD) // Egyptian Informatics Journal. 2011. № 12. P. 37–43.

17. http://www.sakhr.com/ocr.aspx (дата обращения: 08.07.2014).

Статья поступила в редакцию 30 июля 2014 г.

Контактная информация

Редькин Олег Иванович — доктор филологических наук, профессор; oleg_redkin@mail.ru

Берникова Ольга Александровна — кандидат филологических наук, доцент; bernikova@mail.ru

Redkin Oleg I. — Doctor of Philology, Professor; oleg_redkin@mail.ru

Bernikova Olga A. — Candidate of Philology, Associate Professor; bernikova@mail.ru