

**Гернович Т.Д.**  
**ИСТОРИЧЕСКИЕ ИСТОЧНИКИ В ЦИФРОВОЙ СРЕДЕ:**  
**СОВРЕМЕННЫЕ ПОДХОДЫ К АРХЕОГРАФИЧЕСКОЙ ПУБЛИКАЦИИ**

*Ключевые слова: цифровая история, документальное наследие, цифровая публикация документа, археографическая цифровая публикация, цифровой архив.*

В последние десятилетия, в связи развитием технологий оцифровки и размещением цифровых копий исторических источников в интернет, наблюдается увеличение количества публикаций документов. Причем, один и тот же исторический источник может выступать в качестве объекта публикации не только для исследования истории и вспомогательных исторических дисциплин, но и различных гуманитарных наук: филологии, литературоведения, исторической лингвистики, истории искусства, философии, музыковедения, археологии, архивоведения, документоведения, библиотековедения, кодикологии, музеологии и др. Одна и та же публикация исторического памятника используется различными науками в соответствии со сложившимися в рамках каждой из них методами и традициями. Такое сближение гуманитарных наук вокруг общего объекта публикации ставит новые вызовы перед составителем и заставляет переосмыслить прежние методы и принципы. Цифровая научная публикация документа требует много научных сил и технических средств, поэтому должна быть универсальной, междисциплинарной, охватывать различные научные направления и подходы [1; 2].

Теоретические подходы по подготовке научных публикаций документов вырабатывались на протяжении предыдущих столетий и привели к оформлению самостоятельной области исследований, располагающей обширной теоретической литературой, специальными методами, профессиональными ассоциациями, конференциями, журналами и даже специализированными образовательными программами на уровне магистратуры. В Европе дисциплина носит названия Editorik или Editionswissenschaft (в немецком), ecdotica, ecdotique или ecdotics (в итальянском, французском и английском соответственно), в русскоязычной традиции – археография и текстология.

Научный подход при подготовке публикации документа предполагает применение критического метода, поэтому часто издание литературных текстов в текстологии называют «критическими». В отечественной исторической науке такие публикации принято называть «археографическими», в связи с тем, что научные принципы публикации исторических текстов разрабатывались историками в рамках археографии. Вне зависимости от того готовится публикация документа текстологом или археографом, значительная часть работы по подготовке публикации связана с составлением научного аппарата, который позволяет читателю реконструировать происхождение текста и представить содержание в связи с историческим контекстом.

Первые публикации документальных памятников в интернете, которые появились в конце 1990-х – начале 2000-х гг., предлагали пользователям лишь ограниченные возможности для навигации и работы с материалами. Эти инициативы, как правило, носили локальный характер и реализовывались отдельными архивами, библиотеками или университетами. На этом этапе научный аппарат представлялся в виде базы данных, которая объединяла текстовые описания документов с их изображениями, создавая тем самым основу для будущих цифровых коллекций. Однако пользователи таких проектов не могли масштабировать изображения, выполнять поиск по тексту или анализировать содержимое с использованием цифровых инструментов [4].

Примером подобных инициатив могут служить первые цифровые коллекции Библиотеки Конгресса США, которые предоставили доступ к ключевым историческим материалам, или проекты по оцифровке архивов, реализованные на европейских библиотечных порталах. Эти начинания положили основу для дальнейшего развития визуализации документальных памятников, и привели в итоге к двум сложившимся подходам по отображению документа: популярно-ознакомительному и археографическому (научно-критическому).

Первый – популярно-ознакомительный – характерен для библиотечной сферы и предлагает пользователям цифровые изображения документов высокого качества, которые могут сопровождаться информационными материалами об истории документального памятника, или коллекции цифровых изображений с метаданными и возможностью поиска в ней информации по заданным пользователем критериям.

Примером могут служить: Europeana, панъевропейский цифровой архив, предоставляющий доступ к миллионам исторических документов, артефактов, фотографий и аудиовизуальных материалов<sup>1</sup>; Gallica, цифровая библиотека Национальной библиотеки Франции (BNF), включающая книги, рукописи, гравюры и исторические архивы<sup>2</sup>; Deutsche Digitale Bibliothek (DDB), немецкая цифровая библиотека публикующая цифровые копии, хранящиеся в архивах, библиотеках, музеях Германии<sup>3</sup>; Polona, один из крупнейших цифровых архивов Польши, созданный Национальной библиотекой Польши<sup>4</sup>; DIGAR, цифровой архив Национальной библиотеки Эстонии, содержащий старинные книги, периодику и фотографии<sup>5</sup>; ePaveldas (Электронное наследие), национальный цифровой проект Литвы, посвящённый сохранению и публикации культурного и документального наследия<sup>6</sup>; Отдел рукописей Российской национальной библиотеки (РНБ), проект РНБ, в котором представлены факсимиле, автографы рукописей (Остромирово Евангелие, Лаврентьевская летопись, Русская правда и др.), а также справочные материалы о документальных памятниках РНБ<sup>7</sup> и др.

Похожую роль выполняют публикации цифровых копий документов на сайтах архивных учреждений. Так, например, Национальный архив США (National Archives and Records Administration (NARA) (США)) предоставляет доступ к ключевым документам американской истории, включая Конституцию, Декларацию независимости и др.<sup>8</sup>, Литовский государственный исторический архив (Lietuvos valstybės istorijos archyvas) публикует метрические книги, документы XIX–XX вв. и материалы о литовской диаспоре<sup>9</sup>, Российский государственный архив древних актов (РГАДА) представляет на своих страницах цифровые копии грамот, исторических хроник и актов Московской Руси<sup>10</sup>, Национальный цифровой архив Польши (Narodowe Archiwum Cyfrowe, NAC) содержит цифровую коллекцию фотографий, карт, аудиовизуальных записей<sup>11</sup>.

Такие проекты могут быть как простыми, так и интерактивными, однако они все ставят своей целью лишь ознакомление с документальным памятником, публикуя

<sup>1</sup> Europeana. – URL: [europeana.eu](http://europeana.eu) (дата обращения: 22.11.2024).

<sup>2</sup> Gallica. – URL: [gallica.bnf.fr](http://gallica.bnf.fr) (дата обращения: 22.11.2024).

<sup>3</sup> Deutsche Digitale Bibliothek (DDB). – URL: [www.deutsche-digitale-bibliothek.de](http://www.deutsche-digitale-bibliothek.de) (дата обращения: 22.11.2024).

<sup>4</sup> Polona. – URL: [www.polona.pl](http://www.polona.pl) (дата обращения: 22.11.2024).

<sup>5</sup> DIGAR. – URL: [www.digar.ee](http://www.digar.ee) (дата обращения: 22.11.2024).

<sup>6</sup> ePaveldas (Электронное наследие). – URL: [www.epaveldas.lt](http://www.epaveldas.lt) (дата обращения: 22.11.2024).

<sup>7</sup> Отдел рукописей Российской национальной библиотеки (РНБ). – URL: <https://nlr.ru/manuscripts> (дата обращения: 22.11.2024).

<sup>8</sup> National Archives and Records Administration. – URL: <https://www.archives.gov/founding-docs> (дата обращения: 22.11.2024).

<sup>9</sup> Lietuvos valstybės istorijos archyvas. – URL: [archyvai.lt](http://archyvai.lt) (дата обращения: 22.11.2024).

<sup>10</sup> Российский государственный архив древних актов. – URL: [rgada.info](http://rgada.info) (дата обращения: 22.11.2024).

<sup>11</sup> Narodowe Archiwum Cyfrowe. – URL: [nac.gov.pl](http://nac.gov.pl) (дата обращения: 22.11.2024).

исключительно его изображение и не сопровождая его текст информацией, позволяющей погрузиться в исторический и культурный контекст его создания. Для неспециалиста в этом случае документ становится музейным артефактом, который иллюстрирует культурно-исторический контекст, но остается непонятен как по содержанию, так и по его месту в истории.

Так, например, в 2009 г. цифровая публикация фрагмента Синайского кодекса, хранящегося в Британской библиотеке, которая была размещена на ее официальном сайте, привлекла в течении нескольких дней миллион просмотров пользователей. Однако ни один из них впоследствии не подготовил научного исследования. В то же время за предыдущие 25 лет только четыре исследователя смогли познакомиться с этим источником «в живую», все они были признанными специалистами в этом вопросе, и все написали впоследствии научные труды [7].

На платформе International Image Interoperability Framework (IIIF) размещены ссылки на публикации оцифрованных рукописей, хранящихся в крупнейших библиотеках и архивах мира, включая Британскую библиотеку, Ватиканскую библиотеку, Национальную библиотеку Франции и многие другие. На странице Европейской ассоциации цифровых гуманитарных наук (EADH), которая входит в Альянс организаций цифровых гуманитарных наук (ADHO) представлен каталог Каталог цифровых изданий<sup>1</sup>.

На глубокое научное изучение направлены проекты второго типа – археографические (научно-критические) цифровые публикации (critical digital edition, scholarly digital editions). Они размещают дополнительные научные комментарии, аннотации специалистов и справочные материалы. Первые подобные публикации были основаны на цифровых копиях традиционных научных публикаций, которые были выпущены ранее в бумажной форме. Их тексты переводились в цифровой вид и сопровождалась цифровыми изображениями документов и гиперссылками на комментарии и научно-справочный аппарат. В изданиях этого типа оцифрованный текст документа транскрибируется и размечается и в его структуре выделяются не только отдельные элементы (географические объекты, персоналии, исторические понятия и др.), но и осуществляется связь через гиперссылки с научно-критическим аппаратом (комментариями, библиографическими ссылками и др.) [6].

В конце 1980-х гг. как ответ на растущую необходимость систематического подхода к цифровому представлению и анализу текстов в гуманитарных науках был создан Text Encoding Initiative (TEI) – международный консорциум, разрабатывающий стандарты и рекомендации для цифровой разметки текстов в гуманитарных и социальных науках. Основная цель TEI – обеспечить единый формат кодирования текстовых данных, чтобы их можно было легко анализировать, хранить и обмениваться ими. Рекомендации TEI используют XML (eXtensible Markup Language), что делает их совместимыми с современными технологиями и инструментами. TEI позволяет описывать структуру текста (главы, абзацы, стихи), метаданные (автор, дата, источник), лингвистические особенности, комментарии и даже нестандартные элементы, что делает его незаменимым при создании цифровых проектов на основе исторических источников. Так, например, при оцифровке древней рукописи TEI помогает закодировать: структуру текста (например, разделы, строки), палеографические особенности (сокращения, повреждения), метаданные (автор, датировка, место создания) и др.<sup>2</sup>.

---

<sup>1</sup> International Image Interoperability Framework (IIIF). – URL: [https://iiif.io/guides/finding\\_resources/](https://iiif.io/guides/finding_resources/) (дата обращения: 22.11.2024).

<sup>2</sup> Официальный сайт TEI. – URL: <https://tei-c.org/>. Содержит полные руководства, примеры документов и инструменты для работы (дата обращения: 22.11.2024).

Последние проекты археографических цифровых изданий документов, как правило при отображении документа используют двухчастное представление документа на экране: когда в одной части располагается факсимильное изображение, а во второй части его транскрипция. Так, например проекты цифровой публикации «Синайский кодекс»<sup>1</sup>, «Лаврентьевская летопись»<sup>2</sup>, «Супральская летопись»<sup>3</sup>, «Берестяные грамоты»<sup>4</sup> и ряд др. выполнены таким способом.

Археографическая цифровая публикация документа требует транскрибирования текста. Автоматическое распознавание рукописных текстов до недавнего времени представляло значительную сложность. Традиционные OCR-системы (Optical Character Recognition) были разработаны только для печатного текста, поэтому исследователи транскрибировали рукописные тексты в ручном режиме.

В связи с этим использовались различные инструменты, облегчающие и ускоряющие самостоятельное транскрибирование текста, такие как T-PEN (Transcription for Paleographical and Editorial Notation), которые предлагали удобный интерфейс и позволяли осуществлять транскрипцию, добавлять примечания и редакторские комментарии<sup>5</sup>.

Для ускорения обработки больших объемов документов создавались коллаборативные проекты, привлекающие добровольцев к ручному транскрибированию. Одним из таких проектов является «Гражданский архивист» (Citizen Archivist) Национальных архивов США, который сочетает краудсорсинг и современные технологии для распознавания, транскрипции и анализа исторических рукописей, связанных с Гражданской войной в США<sup>6</sup>. Другим купным проектом подобного рода является «Europeana Transcribathon», который предлагает пользователям возможность принять участие в процессе транскрипции рукописных документов, используя онлайн-инструменты платформы Europeana<sup>7</sup>. Участие добровольцев в таких проектах позволяет не только ускорить процесс транскрибирования рукописного текста, но и привлечь внимание общественности к сохранению исторического наследия.

Тенденцией последних лет стало быстрое развитие технологий автоматического распознавания рукописного текста. Современные технологии Handwritten Text Recognition (HTR) активно развиваются благодаря достижениям в области искусственного интеллекта, машинного обучения и компьютерного зрения. Современные системы используют глубокое обучение (Deep Learning), в частности: сверточные нейронные сети (CNN) для извлечения признаков из изображений, рекуррентные нейронные сети (RNN) и их улучшенные версии (LSTM, GRU) для анализа последовательностей символов. Технология HTR создает возможность автоматического распознавания текста на разных языках, включая латиницу, кириллицу, арабскую вязь, китайские иероглифы и другие системы письма. Для повышения точности распознавания HTR-модели обучаются на больших наборах данных, таких как IAM Handwriting Database (английский язык), RIMES (французский язык) или CASIA-HWDB (китайский язык).

Однако для повышения точности распознавания кириллических текстов требуется проводить работу по дополнительному обучению моделей распознавания. Так,

---

<sup>1</sup> Codex Sinaiticus. – URL: [codexsinaiticus.org](http://codexsinaiticus.org) (дата обращения: 22.11.2024).

<sup>2</sup> Лаврентьевская летопись. 1377 г. – URL: [https://expositions.nlr.ru/LaurentianCodex/\\_Project/page\\_Show.php](https://expositions.nlr.ru/LaurentianCodex/_Project/page_Show.php) – (дата обращения: 22.11.2024).

<sup>3</sup> Codex Suprasliensis. – URL: <http://suprasliensis.obdurodon.org/> (дата обращения: 22.11.2024).

<sup>4</sup> Древнерусские берестяные грамоты. – URL: <http://gramoty.ru/birchbark/about-site/> (дата обращения: 22.11.2024).

<sup>5</sup> Сайт проекта T-PEN (Transcription for Paleographical and Editorial Notation). – URL: <https://t-pen.org/TPEN/> – (дата обращения: 22.11.2024).

<sup>6</sup> U.S. National Archives and Records Administration. – URL: <https://www.archives.gov/civil-archivist> (дата обращения: 22.11.2024).

<sup>7</sup> Europeana Transcribe. – URL: <https://europeana.transcribathon.eu/> (дата обращения: 22.11.2024).

например, в проекте «Цифровой Петр» для работы с почерком Петра I, известным своей сложностью и неоднозначностью, для повышения точности распознавания использовался обширный корпус оцифрованных рукописей императора<sup>1</sup>. После завершения этапа обучения благодаря технологиям компьютерного зрения, обработки естественного языка и графы знаний разработчикам проекта удалось добиться качества расшифровки в 97 %, при этом время распознавания составляет 35 страниц в минуту, тогда как у историков на расшифровку каждой страницы столь сложного почерка уходит по несколько часов [3].

Системы распознавания рукописного текста (HTR) в последние годы значительно продвинулись вперед, предлагая широкий спектр инструментов для работы с историческими документами. Одной из наиболее заметных является Transkribus – платформа, специально разработанная консорциумом READ-COOP SCE для распознавания старинных рукописей. Она объединяет передовые технологии искусственного интеллекта (включая нейронные сети и машинное обучение) с инструментами для ручной коррекции, что позволяет работать с рукописными текстами, старопечатными книгами и даже сложными документами с повреждениями или нестандартным форматированием. Transkribus поддерживает множество языков и шрифтов, включая латиницу, кириллицу и исторические алфавиты, а также предоставляет доступ к предобученным моделям<sup>2</sup>.

С начала 2000-х гг. в сфере археографических цифровых публикаций накоплен значительный опыт. Тенденцией последних лет стало появление ресурсов, публикующих ссылки на наиболее значительные проекты в этой области.

Значительный интерес представляет исследовательский опыт Института документологии и научного редактирования (IDE), объединяющего исследователей цифровых гуманитарных наук, занимающихся применением инновационных методов к историческим источникам.

На основе проектной деятельности его исследователи разрабатывают также теоретические вопросы. Результаты исследований были представлены в коллективной монографии «Цифровое научное редактирование: теории и практики», подготовленной Мэтью Джеймсом Дрисколлом и Еленой Пьераццо в 2016 г., где они также попытались определить теоретические рамки понятия «цифровая археографическая публикация» и выделить терминологический аппарат, отражающий специфику этого процесса [5]. Авторы предложили использовать в качестве основного термина – *scholarly digital editions* (цифровые научные публикации).

Под руководством одного из ведущих исследователей Института документологии и научного редактирования (IDE) Патрика Сале (Patrick Sahle) ведётся работа над каталогом Digital Scholarly Editions (последняя версия – v 4.0, 2020ff)<sup>3</sup>.

Как отмечает Патрик Сале, работа над каталогом потребовала выработки ряда критериев, которые позволили бы точно определить, какие цифровые публикации документа стоит относить к научным цифровым изданиям, а какие нет. В течение многих лет он пользовался собственным определением «Edition ist die erschließende Wiedergabe historischer Dokumente». На немецком языке оно формулировалось вокруг центрального, но почти непереводаемого слова *erschließen*, которое охватывает любую деятельность, повышающую объём информации об объекте и, соответственно, его доступность и пригодность для использования. В зависимости от контекста его можно перевести как разрабатывать, раскрывать, делать выводы, интерпретировать. Но ни

<sup>1</sup> Цифровой Петр [Электронный ресурс]. Режим доступа: <https://peterscript.historyrussia.org/> (дата обращения: 22.11.2024).

<sup>2</sup> Transkribus. – URL: <https://www.transkribus.org/> (дата обращения: 22.11.2024).

<sup>3</sup> Каталог Digital Scholarly Editions. – URL: <https://www.digitale-edition.de/exist/apps/editions-browser/about.html> – (дата обращения: 22.11.2024).

одно из этих слов не охватывает всей глубины понятия. Поэтому, чтобы передать основную идею своего подхода он стал использовать термин *critical* (критическое).

В итоге Патрик Сале предлагает следующее определение для понятия *цифровая научная публикация (scholarly digital editions)* – это критическое представление исторических документов [5]. В этом определении «представление» он понимает через перекодировку документа, которая осуществляется на визуальном уровне (факсимиле) или на более абстрактном – текстовом (транскрипция). Далее термин «критическое», используемый в определении, он понимает в широком смысле: это не только текстологическая критика, но также историческая, библиографическая, материальная, визуальная и др. Критика – это любые процессы, в которых осуществляется научное и рефлексивное взаимодействие с материалом, раскрывающее его суть. Критическое отношение требуется и для решения вопроса, каким дополнительным материалом и в какой форме следует снабдить редактируемый текст для лучшего понимания и доступности. Используемый в определении термин «документ», исследователем понимается как любой неабстрактный объект, ставший предметом издания, для которого текст выступает как функция. Он предпочитает термин «документ», термину «текст», так как некоторые документы не содержат текст в его привычном понимании, а некоторые публикации концентрируют свое внимание не только на тексте, но и на материальном носителе. Слово «исторические», которое используется в определении, не означает, что речь идет только об истории. Патрик Сале отмечает, что цифровая научная публикация направлена на раскрытие обстоятельств создания, передачи и изменения документов и текстов. Такие издания преодолевают временную дистанцию, историческую разницу, объясняют то, что непонятно современному читателю.

Как нам представляется, предложенное Патриком Сале определение точно отражает суть современных научных цифровых изданий, их стремление к комплексному, многоаспектному подходу, а также обозначает векторы дальнейшего развития.

Таким образом, цифровая археографическая публикация сегодня представляет собой не просто форму репрезентации исторического источника, а полноценную многоуровневую исследовательскую среду, объединяющую факсимиле, транскрипцию, аннотации и критический аппарат. Такие издания требуют междисциплинарного подхода, использования универсальных стандартов (TEI) и критического осмысления материала во всех его аспектах – текстологическом, историческом, визуальном и материальном. Ключевые направления ее дальнейшего развития как нам представляется, будут связаны с интеграцией автоматизированных инструментов анализа и перекодирования текста, расширением междисциплинарного сотрудничества и разработку новых методологических подходов к работе с историческими источниками в цифровой среде.

1. Афиани, В. Ю. Современные проблемы археографии: историографический аспект / В. Ю. Афиани // Отечественные архивы. – 2019. – № 1. – С. 25–31.

2. Афиани, В. Ю. Цифровая археография – новые возможности и новые проблемы / В. Ю. Афиани // Вспомогательные исторические дисциплины в современном научном знании : материалы XXXIII Международной научной конференции, Москва, 9–10 апреля 2020 года / Российский государственный гуманитарный университет, Историко-архивный институт, Высшая школа источниковедения, специальных и вспомогательных исторических дисциплин; Российская академия наук, Институт всеобщей истории. – Москва : Институт всеобщей истории РАН, 2020. – С. 7–9.

3. Базарова, Т. А. Автографы Петра I: чтение технологиями искусственного интеллекта и создание электронного архива / Т. А. Базарова, М. Е. Проскурякова // Историческая информатика. – 2022. – № 4(42). – С. 179–190.

4. Ловцов, А. С. Проблемы археографии в исследованиях российских архивистов: 1991–2019 гг. / А. С. Ловцов // Вестник ВНИИДАД. – 2020. – № 5. – С. 23–35.

5. Digital Scholarly Editing. / M. J. Driscoll, E. Pierazzo (édit.) – Open Book Publishers, 2016. – URL: <https://books.openedition.org/obp/3381> (дата обращения: 17.12.2024).

6. Digital Technology and the Practices of Humanities Research. / Edmond, Jennifer, éditeur. – Open Book Publishers, 2020. – URL: <https://books.openedition.org/obp/11899> (дата обращения: 17.12.2024).

7. Robinson, P. The Digital Revolution in Scholarly Editing. / P. Robinson // Ars Edendi Lecture Series. / B. Crostini, G. Iversen and B. M. Jensen (eds.) – Vol. IV. – P. 181–207.