ОРГАНИЗАЦИЯ, ФОРМИРОВАНИЕ И СОХРАННОСТЬ ФОНДОВ

УДК 025.2 + 004:02
https://doi.org/10.33186/1027-3689-2025-7-102-121

Анализ критериев отбора данных, предназначенных для длительного хранения в научной библиотеке

Е. В. Бескаравайная

Библиотека по естественным наукам РАН,
Москва, Российская Федерация,
elenabesk@gmail.com, https://orcid.org/0000-0003-2617-1249

Аннотация. Объёмы информации настолько выросли, что вопрос её хранения требует тщательного анализа критериев для отделения данных, имеющих сохраняющуюся ценность, от данных, не имеющих дальнейшей ценности; определения периода времени хранения; обеспечения идентификации записей. Цель данной работы – изучение критериев отбора и выделение из них необходимых и достаточных для практического использования в научных библиотеках при оценке и дифференциации поступающей информации. Результатом стало выделение критериев, сопровождавших сбор и оценку научных данных, с учётом временных рамок их создания и разнообразия форматов. Критерии были разделены на группы по принципу использования: для отбора ресурсов, оценки данных и включения в базу для организации хранения. Для каждого критерия составлен обзор процедур, установлена ответственность специалистов по их применению, определены основные критерии, имеющие наибольшее значение для отбора научной литературы (научная/историческая ценность, юридическая оценка) и второстепенные (полнота, дублетность и др.). Были рассмотрены затруднения, возникающие при практическом использовании отдельных критериев, предложены варианты их преодоления.

Выводы исследования могут быть полезны для сотрудников библиотек, архивов и информационных центров при сборе и анализе данных, предназначенных для организации хранения цифровой информации.

Работа выполнена в рамках научно-исследовательской работы БЕН РАН «Совершенствование методов долгосрочного хранения научной информации как основа естественно-научных знаний».

Ключевые слова: цифровая библиотека, критерии оценки данных, отбор данных для хранения, научная библиотека

Для цитирования: Бескаравайная Е. В. Анализ критериев отбора данных, предназначенных для длительного хранения в научной библиотеке // Научные и технические библиотеки. 2025. № 7. С. 102–121. https://doi.org/10.33186/1027-3689-2025-7-102-121

COLLECTION ORGANIZATION, DEVELOPMENT AND PRESERVATION

UDC 025.2 + 004:02
https://doi.org/10.33186/1027-3689-2025-7-102-121

Analyzing criteria for selecting data intended for long-term preservation at the scientific library

Elena V. Beskaravainaya

Library for Natural Sciences, Russian Academy of Sciences, Moscow, Russian Federation, elenabesk@gmail.com, https://orcid.org/0000-0003-2617-1249

Abstract. The spiraling information flows and information preservation demand in-depth analysis of criteria to separate valuable data from those with no value, to determine preservation time, and to identify the entries. The purpose of the study is to explore into the selection criteria and to specify those necessary and sufficient to be used by scientific libraries to evaluate and differentiate incoming information. Based on the analysis, the author specifies the criteria for scientific data acquisition and evaluation in the context of time and format diversity; the criteria were grouped based on utilization: for resource selection, data evaluation and inclusion into the database to be stored. The procedures for each criteria and related librarian competences are reviewed. The author specifies the criteria most relevant to scientific literature selection (scientific/historical value, legal assessment) and the secondary ones (completeness, duplication, etc.). The related libra-

rian competences are identified. The author also examines challenges that hinder application of individual criteria and proposes the solutions.

The study findings can be useful for librarians, archivists, and information specialists to acquire and analyze the data for further digital preservation.

The study is completed within the Library's for Natural Sciences of the Russian Academy of Sciences R&D project "Improvement of methods of long-term preservation of scientific information as the basis of knowledge in the natural sciences".

Keywords: digital library, data evaluation criteria, data selection for preservation, scientific library

Cite: Beskaravainaya E. V. Analyzing criteria for selecting data intended for long-term preservation at the scientific library // Scientific and technical libraries. 2025. No. 7, pp. 102–121. https://doi.org/10.33186/1027-3689-2025-7-102-121

Объём цифрового контента растёт, а его хранение, включая резервное копирование, требует увеличения объёмов памяти и дополнительных расходов. Хранение всех поступающих данных не выгодно как экономически, так и с точки зрения обнаружения полезных сведений при поиске (сопутствующий шум). Поэтому задачей библиотек становится поиск возможностей дифференциации поступающей информации для организации её хранения.

Организация хранения цифровой информации в библиотеках подразумевает археологию уже имеющихся данных, обнаружение, отбор и обработку новых, извлечение из них знаний. Работа с большим объёмом информации уже на начальном этапе создаёт проблемы при выборе соответствующих и значимых данных для определения их на хранение. На практике многие критерии отбора оказываются либо слишком расплывчатыми для прямого принятия решений, либо слишком конкретными для применения в научной библиотеке.

Цель нашего исследования – формирование группы критериев и включение их в практику принятия решений по отбору данных, предназначающихся для длительного хранения. Важно не просто определить критерии, необходимо выработать алгоритм сбора данных и тех-

нологические опции, которые гармонично сосуществовали бы в библиотеке с другими технологиями, были совместимы со сторонними системами, базами данных, определяли задачи для специалистов и учитывали возникающие проблемы.

В качестве источников информации использовались отечественные и зарубежные публикации: до 2000 г. – для изучения принципов формирования традиционных фондов; с 2015 г. до настоящего времени – для отбора и анализа цифровых данных. В качестве ключевых выражений для поиска использовались: *критерии отбора документов, критерии отбора изданий, оценка научной информации, оценка цифровых данных, сохранение цифровых данных* на русском и английском языках. Для работы были отобраны статьи, описывающие анализ входящего потока информации, либо опыт такой практической работы в научных организациях, библиотеках, аналитических центрах, архивах.

На основе обзора отечественной и зарубежной литературы, её тематического анализа, а также личного опыта и опроса экспертов мы определили основу для выбора критериев по принципу «необходимо и достаточно».

Результаты этого исследования могут использоваться в качестве руководства при работе библиотек с большими данными на этапе сбора информации и организации её длительного хранения.

Заполнение площадей, высокая стоимость поддержания надлежащих условий хранения, обеспечение пожаробезопасности и др. привели к пониманию нецелесообразности хранения всего массива и определили принципы отбора изданий уже на начальном этапе наполнения фонда [**1**]. Универсальными критериями, обосновывающими выбор, являлись соответствие тематике, спрос на издание, экспертная оценка издания учёными и специалистами, библиометрические критерии [**2**] (показатель цитирования, импакт-фактор и др.), стоимость [**3**]. Они и сегодня остаются совокупным набором критериев, на основании которого определяется научная ценность издания независимо от формы его предоставления – бумажной, цифровой или гибридной [**4**, **5**].

Активная цифровизация обусловила включение в фонд научной библиотеки цифровых источников, их структурирование, организацию, определение сроков хранения и ценности, юридическую оценку. Согласно действующему ГОСТу 7.0.102-2018, технологии комплектования библиотечных фондов определены методикой отбора документов

(отбор по формальным и содержательным критериям), выводами на основе экспертных заключений и возможностями комплексных технологий по организации процессов комплектования и по методике отбора документов. В свою очередь расширились требования к критериям, которые должны сегодня не только отвечать всем условиям отбора научной документации, но и быть достаточно универсальными, чтобы применяться к таким форматам входного потока, как гипертексты, аудио- и видеоформаты, материалы экспериментов, наборы данных специализированных баз и др. Поэтому для библиотек и архивов вопрос критериев оценки данных при отборе на долговременное хранение крайне актуален.

Как правило, эксперты [**6**] сходятся на выделении четырёх основных критериев: *научная или историческая ценность*, *уникальность, риск потери*. Далее мнения разделяются, специалисты по цифровым данным предлагают включить в список, кроме перечисленных критериев *новизну источника или типа данных, эксплуатационные преимущества, возможность репликации* [**7**], *невоспроизводимость, полноту документации, завершённость оценки данных* [**8**], *«находимость», доступность, возможность повторного использования* [**9**].

Сложный вопрос выбора цифровых данных для длительного хранения требует тщательного рассмотрения критериев отбора по нескольким направлениям.

1. Критерии для отбора источников информации

На первом этапе необходимо определить систему критериев отбора источников информации [**10**], данные из которых в дальнейшем будут исследованы на предмет организации хранения. Основным критерием для рассмотрения ресурса является его *информативность,* то есть способность предоставлять полезную, точную и актуальную информацию, будь то журнал, интернет-портал, база данных, мультимедийный ресурс или онлайн-коллекция. Для научных ресурсов оценка информативности включает совокупность нескольких специфичных аспектов [**11**]: количественных (цитирование в тематических источниках, импакт-факторы журналов, контент-анализ и др.) и качественных (отзывы и рекомендации авторитетных экспертов в данной области, мнения учёных и др.). Мы определили область оценки в критериях, отобранных под задачи извлечения информации для хранения в научной библиотеке (табл. 1).

Таблица 1

Критерии для оценки ресурсов

Критерии	Область оценки	Ответственный
Соответствие тематике (научная ценность)	Изучить соответствие содержания ресурса тематике (особенно важно для мультидисциплинарных и новых ресурсов); рассмотреть свидетельства упоминания ресурса в публикациях или отчётах, представляющих дисциплину данных	Библиографы, эксперты, учёные
Актуальность ресурса	Сопоставить соответствие информации на представленном ресурсе текущим исследованиям и тенденциям в данной области, а также проанализировать частоту обновлений для включения новых данных и исследований	Библиографы, эксперты, учёные
Релевантность	Провести опрос, насколько информация на ресурсе соответствует потребностям и интересам пользователей, подтверждена ли другими источниками или исследованиями	Эксперты, учёные
Репутация	Изучить текущую информацию о цитировании из источников научного и образовательного характера, опубликованных в рецензируемых научных журналах или отчётах, полученных от учёных, представляющих данную дисциплину; для изданий – проверить рецензирование, присутствие авторитетных рецензентов	Библиографы
Законность и конфиденциальность ресурса-кандидата	Проверить ресурс на содержание информации, распространение которой запрещено законом (дискриминация, политические лозунги и др.) и ограничения (авторское право, личные данные)	Юрисконсульт
Стабильность	Оценить время существования ресурса, полноту выпусков, наличие недостающих выпусков (для электронных изданий)	Библиографы
Доступность	Выяснить, насколько легко пользователи могут получить доступ к ресурсу и использовать его	Сотрудник IT-отдела
Наглядность	Определить чёткость и доходчивость изложения информация на сайте; наличие интерактивных элементов, графиков, таблиц, видео	Учёные, сотрудник IT-отдела
Технологическая доступность	Проверить формат ресурса на соответствие техническим критериям долгосрочного хранения: полноту гиперссылок, мультимедиа, документации, необходимых для облегчения будущего обнаружения, доступа и использования	Сотрудник IT-отдела

Основным ресурсом для наших целей остаются тематические базы данных и специализированные поисковые системы [**12**, **13**] по физико-химической биологии, математике, физике, астрономии и др., такие как: Medline (https://pubmed.ncbi.nlm.nih.gov/), Medlib (https://www.medlib.ru/), Math-Net (https://www.mathnet.ru/), MathSciNet (https://mathscinet.ams.org/mathscinet), zbMATH (https://zbmath.org/), MatWeb (https://www.matweb.com/), Math.ru (ww.math.ru), GeoNames (www.geonames.org), MedBioWorld (https://www.medbioworld.com/), AGRIS (https://www.fao.org/agris/ru), Worldmapper (worldmapper.org), ExPASy (www.expasy.org), KEGG (https://www.genome.jp/kegg/), NCBI databases (www.ncbi.nlm.nih.gov), Pathway Commons (pathwaycommons.org), Protein Information Resource (proteininformationresource.org/), UniProt (www.uniprot.org), ZOOINT (zin.ru/projects/zooint_r/index.html) и др.

Новый ресурс попадает в поле зрения, как правило, по рекомендации учёных, но на поверку часто не проходит тест на дублетность при сравнении с проверенными ресурсами. Тем не менее, если он соответствует критериям отбора, то остаётся в работе, и в дальнейшем исследуется на наличие уникальной информации.

Одним из важнейших элементов развития фондов научной библиотеки остаётся централизованная подписка на базы данных, обеспечивающая доступ к актуальным и проверенным информационным ресурсам. Среди преимуществ, включающих экономическую выгоду от пользования единой подпиской, присутствуют неявные, но очень важные стратегии информационного обслуживания в научных библиотеках, например, единая платформа доступа ко множеству ресурсов; возможность работы с редкими/специализированными источниками, недоступными для отдельных пользователей или учреждений; контроль за правами доступа и соблюдением лицензионных соглашений и др. К сожалению, недостаток финансирования ограничивает возможность подписки на все необходимые ресурсы, а стоимость подписки на научные базы данных сильно варьируется из года в год. Участие в библиотечных консорциумах (АРБИКОН, НЭИКОН, Консорциум российских научных библиотек) для совместного приобретения подписок уже много лет становится выходом из сложной экономической ситуации.

Отметим возникающие при отборе ресурсов трудности: дублирование информации на различных ресурсах, сложность привлечения экспертов, изменение финансовой политики держателей ресурсов с огра-

ничением доступа, устаревание данных. Отметим, что ресурс – это площадка для сбора данных. Для дальнейшей работы она должна отвечать основным критериям: соответствие тематике (научная ценность), актуальность, релевантность, законность и конфиденциальность, и, по возможности, отвечать условиям стабильности, доступности и наглядности.

2. Критерии отбора данных (оценка)

Исходя из того, что деятельность по поиску, отбору и сохранению информации регламентируется разными нормативными актами и исполняется сотрудниками разных служб, следует разделять критерии для отбора данных на хранение и критерии для включения их в базу. В первом случае работу выполняют библиографы и специалисты информационно-аналитических служб, их квалификация должна позволять выделить данные по тематике, провести экспертизу ценности документов, библиометрический анализ, определить репутацию издания. Для обеспечения идентификации записей, проверки их полноты и принятия мер по их сохранению необходимы навыки IT-специалистов. Такой сложный процесс, как отбор и оценка видео- и аудиоформатов, потребует последовательной работы библиографов, IT-специалистов, иногда юрисконсультов. А решение по таким критериям, как, например, *приоритетные данные* или *потенциальное использование,* невозможно без привлечения экспертного мнения.

Методология отбора данных может варьироваться в зависимости от типа учреждения и его специфических задач; применение критериев, представленных в табл. 2, поможет организовать отбор, соответствующий миссии и политике научной библиотеки.

Таблица 2

Критерии отбора информации, предназначенной для хранения в библиотеке

Критерии	Область проверки
Научная ценность/ тематика	Изучить текущие свидетельства цитирования (исследовательское или образовательное использование), опубликованные в рецензируемых научных публикациях или отчётах, полученных от признанного сообщества учёных, представляющих данную дисциплину; проверить, являются ли наборы данных новыми или уникальными

Продолжение таблицы 2

Критерии	Область проверки
Актуальность данных	Сопоставить, насколько информация соответствует текущим исследованиям и тенденциям в данной области (цитат-анализ, быстроцитируемые работы, поддержка грантами)
Полнота	Проверить, насколько данные охватывают все возможные сценарии для выполнения определённой задачи или включения в дальнейший анализ
Целостность	Проверить, не включают ли данные пропуски или недостающие значения; целесообразно одновременно подвергать оценке не отдельную запись, а идентичные документы из различных источников
Дублетность	Необходимо выявить документы, в которых присутствует повторение информации (например, копирование, тиражирование, частичное воспроизведение), оставив наиболее полный документ; при оценке дубликатов необходимо учитывать *подлинность*, копий – сохранность и способность выдержать длительные сроки хранения. При переводе в цифровую среду бумажных документов среди идентичных вариантов учитываются *физическое состояние* и степень *сохранности* (чем меньше сохранился подлинный документ, тем ценнее он для определения на хранение)
Историческая ценность	Критерий отрабатывается в двух направлениях. *Ценность происхождения*: организационный и функциональный контекст, в котором они были созданы, и *ценность содержания* (например, фиксирование события в момент, когда оно происходит, видеозапись выступления)
Уникальность данных (риск потери)	Проверить, будет ли потеряна информация, если данные не сохранены (то есть потеря сведений о фактах, которые никогда не повторятся, например, геоданные или фото с археологических раскопок); обратить внимание на наличие факторов (палеографических, художественных и др.), повышающих ценность документов
Потенциальное использование и применение	Оценка включает в себя вывод о предполагаемом будущем использовании на основе данных текущей исследовательской и образовательной ценности, например, потенциальное участие их в статистике/анализе. Необходимо обосновать, оправдывает ли потенциал будущего использования данных затраты на их долгосрочное архивирование (эту информацию можно получить из ссылок на публикации, в которых использовались данные, или из авторитетных источников), проверить не являются ли они частью набора данных, участвующих в комплексном анализе

Окончание таблицы 2

Критерии	Область проверки
Невоспроизводимость (стоимость замещения)	Проверить, насколько воспроизводство данных невозможно или чрезмерно дорогостояще (отсутствие достаточных контекстных материалов для повторного использования, значительные ресурсы для преобразования, или ресурсы, которые невозможно восстановить)
Юридическая оценка	Проверить, содержит ли документ конфиденциальную информацию или информацию, разглашающую личности отдельных людей, не нарушает ли авторское право (незаконное использование, копирование и распространение произведения); требуется ли разрешение правообладателя на весь документ или его части (в том числе для видео- и аудиоконтента). Необходимо проверять временные рамки для авторского права на произведения и патенты

Разработка иерархии критериев при оценке и выборе данных – процесс сложный и неоднозначный. Часто возникает ситуация, когда документ соответствует приоритетному критерию *подлинности,* но не проходит по критерию *полноты* (текст в нём не читается или утрачен). В этом случае на длительное хранение направляется более качественная копия после проверки наличия подписей и печатей. Таким образом, выбор данных происходит на основе: а) значимости критериев; б) оценки документов.

В литературе, рассматривающей оценку документов, предназначенных для хранения, нам встретились различные принципы ранжирования критериев [**14**]. На практике чаще всего используются три основные методики [**15**, **16**]:

метод одиночного критерия: каждому из критериев на основе субъективной оценки присваивается балл, а затем выводится средний балл для отдельного критерия;

метод парного сравнения: сравниваются два критерия и определяется приоритетный по мнению респондента [**17**]; следует заметить, что данный метод может быть использован для сравнения таких сложных категорий, как качество изображений и видео [**18**];

метод экспертных суждений [**19**]: незаменим при анализе сложных проблем или оценке новых технологий, для которых не установлены контрольные показатели или отсутствуют объективные данные; вы-

воды делаются путём объединения экспертных суждений как совокупности количественных оценок.

Очевидно, что нельзя принимать решение по единому критерию, отсюда сложность создания их иерархии. Приоритетность часто зависит от специфики библиотек [**20**]: для исторических библиотек важны критерии подлинности, времени и места создания, позволяющие оценивать документ как свидетельство о существенных событиях [**21**]; для национальных библиотек в первую очередь ценны документы, представляющие интерес для фольклористов, языковедов, краеведов [**22**, **23**], для традиционных библиотек – документы, обладающие научной и художественной ценностью, уникальностью, сохранностью издания [**24**], а для научных библиотек первостепенными являются научная значимость и потенциальное использование [**25**].

3. Критерии для включения данных в базу

Итак, документы прошли первичный отбор и были определены на хранение. Они существенно различаются по типу данных (экспериментальные, данные наблюдений, вторичные, справочные, коллекции, изображения, видео, программное обеспечение для визуализации специфических данных или создания моделей, патенты и др.), по уровню внутренней организации (обработанные, необработанные, вспомогательные), по происхождению (результат исследования, побочный продукт исследования, рабочий процесс и др.). Более того, собранные данные относятся к разным технологическим периодам, а, следовательно, отличаются форматом. Анализ данных для включения в базу хранения требует совокупного учёта критериев, обеспечивающего не только сохранность информации, но и её целостность, безопасность и эффективность дальнейшего использования (табл. 3).

Таблица 3

Критерии для включения данных в базу хранения

Критерии для отбора	Область проверки
Законность и конфиденциальность	Проверить соблюдение правовых норм и защиту личных данных при сборе, хранении и обработке; ограничения или требования по сохранению и будущему доступу. Возможно, следует получить явное конкретное и добровольное согласие пользователя на сбор и обработку материалов или своих данных

Окончание таблицы 3

Критерии для отбора	Область проверки
Качество данных	Данные должны быть точными и полными, проверяться на ошибки и наличие опечаток, неверных значений, дублирования и др.
Технологическая осуществимость	Проверить формат на соответствие техническим критериям уровня обслуживания долгосрочного архива и программному обеспечению для доступа и просмотра; при возникновении технологических проблем следует использовать записи, содержащие похожую документацию или другой формат
Полнота данных	Все необходимые поля данных должны быть заполнены, недостающие данные могут привести к неполному анализу или неправильным выводам
Согласованность	Данные должны быть согласованы между собой, если они хранятся в таблицах, таблицы должны быть везде одинаковыми
Достоверность	Данные должны соответствовать определённым правилам или форматам, например, даты – в правильном формате, а числовые значения – в допустимом диапазоне
Уникальность	Каждая запись должна быть уникальной во избежание дублирования данных, что может быть достигнуто с помощью уникальных идентификаторов
Соблюдение норм и стандартов	Данные должны соответствовать регуляторным требованиям, таким как GDPR, HIPAA и др., что может повлечь необходимость их хранения в определённых условиях
Происхождение	Предусмотреть возможность отследить происхождение данных, чтобы понять, откуда они были получены и как обработаны (ссылки на данные будут большим показателем их ценности, чем цитаты)
Производительность	Организация данных должна обеспечить высокую производительность при выполнении запросов и операций
Удобство использования	Данные важно сохранять в формате, позволяющем другим использовать его без затрат или других ограничений; необходимо проверить, достаточно ли метаданных и документации для того, чтобы набор данных можно было легко использовать и понимать вне его исходного контекста

Важнейшими критериями отбора данных на включение в базу являются *законность, конфиденциальность, соблюдение норм и стандартов*. Только после установления соответствия этим критериям документы могут быть оценены на *полноту, качество, уникальность* и др.

На практике приём данных реализуется в виде пакетов (например, ETL), эффективных с точки зрения масштабирования, сохранности и точного анализа, потока (например, Apache Kafka, Spark Streaming, Elasticsearch и др.), способного анализировать большие данные в реальном времени, либо гибридов (например, Apache Flink. Lambda, Kappa), сочетающих обе парадигмы в одном решении. Выбор технологии приёма данных зависит от их типов и объёмов, совместимости с установленными системами и наличием квалифицированной поддержки.

Основными задачами, связанными с автоматизированным приёмом данных, являются их очистка, интеграция, синхронизация, проверка дублирования и пропущенных значений.

В литературе, кроме упомянутых, нам встретились такие критерии, как удобство использования набора данных, полнота метаданных, организация доступа, шифрование данных, производительность базы, масштабируемость, резервное копирование и др., которые мы сознательно не рассматриваем, так как считаем, что эти критерии описывают организацию хранения, а не оценку данных.

Чтобы раскрыть потенциальные проблемы, вытекающие из процесса отбора и оценки данных, необходимо участвовать в практических экспериментах. Хочется сказать несколько слов о затруднениях, возникших перед сотрудниками нашей библиотеки на этапе анализа/отбора цифровой информации для организации её на хранение, и возможных путях их решения:

1. Определение круга ядерных изданий по тематике по таким показателям, как репутация издания, цитирование и др., – обычная работа библиографа или сотрудника информационно-аналитического отдела. Однако сегодня появилась масса онлайн-изданий, срок жизни которых не позволяет сделать выводы о цитируемости, импакт-факторе или репутации ресурса. Возможное решение – библиометрический анализ авторов из таких изданий, поиск и привлечение экспертов для оценки наличия принципиально новых направлений.

2. Отбор видео- и аудиоформатов требует оборудования и специальной подготовки для определения качества изображения и звука, совместимости с устройствами, эффективности сжатия и иных характеристик. Возможное решение – применение метода парного сравнения.

3. Один из самых сложных вопросов – оценка потенциального использования. Данные, которые могут быть неактуальны для первоначальной цели, станут ценными для исторического исследования или статистического анализа. Возможное решение – сохранение данных, по которым нет однозначного вывода, путём архивирования.

4. Собранные данные становятся менее актуальными из-за эволюции современных систем, прекращения поддержки, изменения требований пользователей, физической совместимости с оборудованием и др. Возможное решение – конвертация данных из устаревших форматов (миграция) и сохранение в открытых форматах (PDF, CSV, XML, JSON), менее подверженных устареванию по сравнению с проприетарными форматами уже на стадии организации данных на хранение.

5. Идентификация похожей информации из различных источников и выявление повторяемости части информации в других требует значительных усилий на очистку и подготовку данных для анализа (разбивка текста на токены, удаление стоп-слов и др.). Возможные решения – применение методов машинного анализа (косинусное сходство и TF-IDF) – для текстов; кластеризация и корреляционный анализ для численных данных; свёрточные нейронные сети и методы хеширования (pHash или dHash) для быстрого сравнения изображений; динамическое временное выравнивание (DTW) и автокорреляция – для выявления повторяющихся паттернов в данных.

6. Оценка данных наблюдений является сложной задачей, поскольку они состоят из первичных и вторичных данных, каждые из которых, в свою очередь, включают совокупность необработанных и обработанных данных. Возможное решение: для вторичных данных сохранять только обработанные варианты, для первичных – всю информацию.

Выводы

Данные на хранение в научной библиотеке принимаются только после тщательного анализа по совокупности критериев. Если подтверждена их научная или историческая ценность (данные, имеющие отношение к приоритетным исследованиям, национальные коллекции, данные высокого спроса), они поступают на хранение, как бы трудозатратна ни была их обработка. Напротив, если данные «не прошли» по критерию ценности, они, скорее всего, не будут оцениваться и по таким

критериям, как потребности пользователей, аналитическая ценность, удобство, доступность и др. Таким образом, основным критерием является *научная или историческая ценность*, затем следует *юридическая оценка*, остальные критерии вспомогательные и, как правило, применяются совокупно (см. рисунок).

Поэтапная работа с критериями по отбору данных, предназначенных на хранение

Работа по отбору и анализу данных с использованием изложенных критериев проводится в БЕН РАН. На сегодняшний день имеется два типа данных:

1. Данные, соответствующие разработанным критериям и отобранные для долгосрочного хранения. Собранные из авторитетных источников и соответствующие критериям оценки, они имеют долгосрочный потенциал хранения и вторичного анализа.

2. Необработанные данные (шумные, неполные данные, промежуточные результаты исследований, ранние резервные копии и др.) после обработки направят на длительное хранение или удалят.

Приём данных требует *планирования и организации*, задействуя время, людей, оборудование, инфраструктуру, технологии для анализа. И это только на начальном этапе, до организации хранения и предоставления доступа к данным. Так, сортировка входящего потока с гетерогенными данными различных размеров, типов и форматов пока не осуществляется без участия сотрудников с привлечением программных инструментов; для обработки шумных и неполных данных на помощь сотрудникам всё чаще приходят методы машинного обучения и интеллектуального анализа; а такие сложные процессы, как обработка динамических данных, требующие глобального пересчёта каждый раз, когда одна из входных схем модифицируется, могут осуществляться только с использованием технологий с высокой пропускной способностью.

Организацию хранения данных в течение длительного времени следует рассматривать ещё по одному критерию – *экономической эффективности*. Увеличение объёма данных потребует дополнительных затрат на более мощное оборудование, программное обеспечение, обслуживание и управление, смену выбранных технологий и поставщиков услуг.

Технологии хранения данных быстро развиваются и систематический анализ затрат поможет определить, насколько жизнеспособно текущее решение. Хочется надеяться, что хорошо продуманные и организованные приём, оценка и анализ данных на начальном этапе помогут съэкономить деньги библиотеки за счёт дальнейшей автоматизации дорогостоящих и трудоёмких процессов.

Список источников

1. **Столяров Ю. Н.** Библиотечный фонд : учебное для студентов вузов, обучающихся по направлению подготовки 071900 «Библиотечно-информационная деятельность»: квалификация бакалавр / Ю. Н. Столяров. Санкт-Петербург : Издательство Профессия, 2015. 383 с. ISBN 978-5-904757-87-8.

2. **Гуреев В. Н.** Использование библиометрии для оценки значимости журналов в научных библиотеках. (Обзор) // Научно-техническая информация. Сер. 1: Организация и методика информационной работы. 2015. № 2. С. 8–19.

3. **Земсков А. И., Евстигнеева Г. А.** Роль библиотек на мировом рынке научных публикаций // Вестник Российского фонда фундаментальных исследований. 2005. №. 4. С. 51–56.

4. **Вихрева Г. М., Подкорытова Н. И.**, **Федотова О. П.** Исследования системы фондов библиотек Сибирского отделения Российской академии наук // Труды ГПНТБ СО РАН. 2021. № 2 (10). С. 23–33. https://doi.org/10.20913/2618-7575-2021-2-23-33.

5. **Вихрева Г. М., Федотова О. П.** Проблемы формирования библиотечного фонда в контексте философии ценностей // Библиосфера. 2017. № 2. С. 12–16. https://doi.org/10.20913/1815-3186-2017-2-12-16.

6. **Whyte A., Wilson A.** How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre. 2010. URL: dcc.ac.uk/resources/how-guides/appraise-select-data (access: 23.02.2025).

7. **Data** management costing tool and checklist, version 3, UK Data Service.: UK Data Service (UKDS). 2015. URL: ukdataservice.ac.uk/manage-data/plan/costing (access: 03.03.2025).

8. **DCC**, Five steps to decide what data to keep: checklist for appraising research data // Edinburgh: Digital Curation Centre. 2014. Vol. 1. URL: dcc.ac.uk/resources/how-guides (access: 11.02.2025).

9. **Tenopir C., Rice N. M., Allard S., Baird L., Borycz J., Christian L. et al.** Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide // PLoS ONE. 2020. Vol. 15. № 3. e0229003. https://doi.org/10.1371/journal. pone.0229003.

10. **Антопольский А. Б.** Будущее научных коммуникаций и научной информации // Информация и инновации. 2019. Т. 14. № 1. С. 7–17.

11. **Кириллова О. В.** Редакционная подготовка научных журналов по международным стандартам: рекомендации эксперта БД Scopus. Москва, 2013. 90 с. ISBN 978-5-518-73515-6.

12. **Мохначёва Ю. В., Цветкова В. А.** Возможные пути получения научной информации в новых условиях // Управление наукой: теория и практика. 2023. Т. 5. № 3. С. 117–158. https://doi.org/10.19181/smtp.2023.5.3.9.

13. **Цветкова В. А., Мохначёва Ю. В., Харыбина Т. Н., Бескаравайная Е. В., Митрошин И. А.** Пространство знаний: подходы к извлечению знаний из научных текстов // Информационные ресурсы России. 2019. № 2 (168). С. 31–34.

14. **Земсков А. И.** Data Curation хранение научных данных и обслуживание ими новое направление деятельности библиотек // Научные и технические библиотеки. 2013. № 2. С. 85–101.

15. **Краснов Ф. В., Дементов А. В., Шварцман М. Е.** Использование тематических моделей для парного сравнения коллекций научных статей // Информатика и её применения. 2020. Т. 14. №. 3. С. 129–135.

16. **Анохин А. М., Глотов В. А., Павельев В. В., Черкашин А. М.** Методы определения коэффициентов важности критериев. Автоматика и телемеханика // Автоматика и телемеханика. 1997. № 8. С. 3–35.

17. **Ozbey C., Dincsoy O.** An Efficient Pairwise Comparison Scheme for Document Ranking. 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 2020. Pp. 1–4. https://doi.org/10.1109/SIU49456.2020.9302078.

18. **Zhang Z., Zhau J., Liu N., Gu, X., Zhang Y.** An improved pairwise comparison scaling method for subjective image quality assessment/2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Cagliari, Italy, 2017. Pp. 1–6. https://doi.org/10.1109/BMSB.2017.7986235.

19. **Хакимова Е. В.** Автоматизированная система комплектования книжного фонда на основе экспертных оценок // Перспективы развития информационных технологий. 2011. № 3–2. С. 258–262.

20. **Гуреев В. Н.** Модели и критерии отбора изданий в фонд научной библиотеки // Научные и технические библиотеки. 2015. № 7. С. 31–50.

21. **Иванов Ю. Н.** Основные критерии отбора документов с позиций исторической ценности и достоверности // Крыніцазнаўства, археаграфія, архівазнаўства ў XX–XXI ст. у Беларусі : зб. навук. артыкулаў, прысвечаных 100-годдзю з дня нараджэння М. М. Улашчыка / рэдкал. : С. М. Ходзін (адк. рэд.) [і інш.]. Мінск : БДУ, 2007. С. 203–210. ISBN 978-985-485-977-4.

22. **Браккер Н. В., Куйбышев Л. А.** Сбор и архивирование сетевых ресурсов. Опыт национальных библиотек зарубежных стран // Библиотековедение. 2013. № 2. С. 88–96. https://doi.org/10.25281/0869-608X-2013-0-2-88-96.

23. **Алебастрова Е. С.** Краеведческий контент Электронной библиотеки Национальной библиотеки Республики Саха (Якутия): проблемы формирования и использования // Вестник национальной библиотеки Республики Саха (Якутия). 2019. № 2 (19). С. 19–24.

24. **Степанов В. К.** Формирование цифровых коллекций в традиционных библиотеках // Научные и технические библиотеки. 2007. № 2. С. 89–94.

25. **Бочарова Е. Н.** Критерии отбора документов в фонд научной библиотеки // Взаимовлияние информационно-библиотечной среды и общественных наук : сборник научных статей / РАН. ИНИОН. Фундаментальная библиотека ; науч. ред. Л. Н. Тихонова, А. А. Джиго. Москва : Институт научной информации по общественным наукам РАН, 2018. С. 62–78.

Reference

1. **Stoliarov Iu. N.** Bibliotechny`i` fond : uchebnoe dlia studentov vuzov, obuchaiushchikhsia po napravleniiu podgotovki 071900 «Bibliotechno-informatcionnaia deiatel`nost`»: kvalifikatciia bakalavr / Iu. N. Stoliarov. Sankt-Peterburg : Izdatel`stvo Professiia, 2015. 383 s. ISBN 978-5-904757-87-8.

2. **Gureev V. N.** Ispol`zovanie bibliometrii dlia ocenki znachimosti zhurnalov v nauchny`kh bibliotekakh. (Obzor) // Nauchno-tekhnicheskaia informatciia. Ser. 1: Organizatciia i metodika informatcionnoi` raboty`. 2015. № 2. S. 8–19.

3. **Zemskov A. I., Evstigneeva G. A.** Rol` bibliotek na mirovom ry`nke nauchny`kh publikatcii` // Vestneyk Rossii`skogo fonda fundamental`ny`kh issledovanii`. 2005. №. 4. S. 51–56.

4. **Vikhreva G. M., Podkory`tova N. I., Fedotova O. P.** Issledovaniia sistemy` fondov bibliotek Sibirskogo otdeleniia Rossii`skoi` akademii nauk // Trudy` GPNTB SO RAN. 2021. № 2 (10). S**.** 23–33. https://doi.org/10.20913/2618-7575-2021-2-23-33.

5. **Vikhreva G. M., Fedotova O. P.** Problemy` formirovaniia bibliotechnogo fonda v kontekste filosofii cennostei` // Bibliosfera. 2017. № 2. S**.** 12–16. https://doi.org/10.20913/1815-3186-2017-2-12-16.

6. **Whyte A., Wilson A.** How to Appraise and Select Research Data for Curation. DCC How-to Guides. Edinburgh: Digital Curation Centre. 2010. URL: dcc.ac.uk/resources/how-guides/appraise-select-data (access: 23.02.2025).

7. **Data** management costing tool and checklist, version 3, UK Data Service.: UK Data Service (UKDS). 2015. URL: ukdataservice.ac.uk/manage-data/plan/costing (access: 03.03.2025).

8. **DCC**, Five steps to decide what data to keep: checklist for appraising research data // Edinburgh: Digital Curation Centre. 2014. Vol. 1. URL: dcc.ac.uk/resources/how-guides (access: 11.02.2025).

9. **Tenopir C., Rice N. M., Allard S., Baird L., Borycz J., Christian L. et al.** Data sharing, management, use, and reuse: Practices and perceptions of scientists worldwide // PLoS ONE. 2020. Vol. 15. № 3. e0229003. https://doi.org/10.1371/journal. pone.0229003.

10. **Antopol`skii` A. B.** Budushchee nauchny`kh kommunikatcii` i nauchnoi` informatcii // Informatciia i innovatcii. 2019. T. 14. № 1. S. 7–17.

11. **Kirillova O. V.** Redaktcionnaia podgotovka nauchny`kh zhurnalov po mezhdunarodny`m standartam: rekomendatcii e`ksperta BD Scopus. Moskva, 2013. 90 s. ISBN 978-5-518-73515-6.

12. **Mokhnachyova Iu. V., Tcvetkova V. A.** Vozmozhny`e puti polucheniia nauchnoi` informatcii v novy`kh usloviiakh // Upravlenie naukoi`: teoriia i praktika. 2023. T. 5. № 3. S. 117–158. https://doi.org/10.19181/smtp.2023.5.3.9.

13. **Tcvetkova V. A., Mokhnachyova Iu. V., Hary`bina T. N., Beskaravai`naia E. V., Mitroshin I. A.** Prostranstvo znanii`: podhody` k izvlecheniiu znanii` iz nauchny`kh tekstov // Informatcionny`e resursy` Rossii. 2019. № 2 (168). S. 31–34.

14. **Zemskov A. I.** Data Curation khranenie nauchny`kh danny`kh i obsluzhivanie imi novoe napravlenie deiatel`nosti bibliotek // Nauchny`e i tekhnicheskie biblioteki. 2013. № 2. S. 85–101.

15. **Krasnov F. V., Dimentov A. V., Shvartcman M. E.** Ispol`zovanie tematicheskikh modelei` dlia parnogo sravneniia kollektcii` nauchny`kh statei` // Informatika i eyo primeneniia. 2020. T. 14. №. 3. S. 129–135.

16. **Anohin A. M., Glotov V. A., Pavel`ev V. V., Cherkashin A. M.** Metody` opredeleniia koe`ffitcientov vazhnosti kriteriev. Avtomatika i telemehanika // Avtomatika i telemehanika. 1997. № 8. S. 3–35.

17. **Ozbey C., Dincsoy O.** An Efficient Pairwise Comparison Scheme for Document Ranking. 2020 28th Signal Processing and Communications Applications Conference (SIU), Gaziantep, Turkey, 2020. Pp. 1–4. https://doi.org/10.1109/SIU49456.2020.9302078.

18. **Zhang Z., Zhau J., Liu N., Gu, X., Zhang Y.** An improved pairwise comparison scaling method for subjective image quality assessment/2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Cagliari, Italy, 2017. Pp. 1–6. https://doi.org/10.1109/BMSB.2017.7986235.

19. **Hakimova E. V.** Avtomatizirovannaia sistema komplektovaniia knizhnogo fonda na osnove e`kspertny`kh ocenok // Perspektivy` razvitiia informatcionny`kh tekhnologii`. 2011. № 3–2. S. 258–262.

20. **Gureev V. N.** Modeli i kriterii otbora izdanii` v fond nauchnoi` biblioteki // Nauchny`e i tekhnicheskie biblioteki. 2015. № 7. S. 31–50.

21. **Ivanov Iu. N.** Osnovny`e kriterii otbora dokumentov s pozitcii` istoricheskoi` cennosti i dostovernosti // Крыніцазнаўства, археаграфія, архівазнаўства ў XX–XXI ст. у Беларусі : зб. навук. артыкулаў, прысвечаных 100-годдзю з дня нараджэння М. М. Улашчыка / рэдкал. : С. М. Ходзін (адк. рэд.) [і інш.]. Мінск : BDU, 2007. S. 203–210. ISBN 978-985-485-977-4.

22. **Brakker N. V., Kui`by`shev L. A.** Sbor i arhivirovanie setevy`kh resursov. Opy`t natcional`ny`kh bibliotek zarubezhny`kh stran // Bibliotekovedenie. 2013. № 2. S. 88–96. https://doi.org/10.25281/0869-608X-2013-0-2-88-96.

23. **Alebastrova E. S.** Kraevedcheskii` kontent E`lektronnoi` biblioteki Natcional`noi` biblioteki Respubliki Saha (Iakutiia): problemy` formirovaniia i ispol`zovaniia // Vestneyk natcional`noi` biblioteki Respubliki Saha (Iakutiia). 2019. № 2 (19). S. 19–24.

24. **Stepanov V. K.** Formirovanie tcifrovy`kh kollektcii` v traditcionny`kh bibliotekakh // Nauchny`e i tekhnicheskie biblioteki. 2007. № 2. C. 89–94.

25. **Bocharova E. N.** Kriterii otbora dokumentov v fond nauchnoi` biblioteki // Vzaimovliianie informatcionno-bibliotechnoi` sredy` i obshchestvenny`kh nauk : sbornik nauchny`kh statei` / RAN. INION. Fundamental`naia biblioteka ; nauch. red. L. N. Tihonova, A. A. Dzhigo. Moskva : Institut nauchnoi` informatcii po obshchestvenny`m naukam RAN, 2018. S. 62–78.

Информация об авторе / Author

Бескаравайная Елена Вячеславовна – старший научный сотрудник Библиотеки по естественным наукам РАН, Москва, Российская Федерация

elenabesk@gmail.com

Elena V. Beskaravainaya – Senior Researcher, Library for Natural Sciences, Russian Academy of Sciences, Moscow, Russian Federation

elenabesk@gmail.com