

РАБОЧАЯ ТЕТРАДЬ ДЛЯ ИТ-ДИРЕКТОРА И ВЛАДЕЛЬЦА ДОКУМЕНТООБОРОТА

Оценка корпоративного архива перед RAG-проектом

Формы инвентаризации, проверка качества сканов, замер точки отсчёта, эталонный набор вопросов и итоговый лист с тремя исходами

Инвентаризация

Проверка сканов

Признаки неготовности

Точка отсчёта

Эталонный набор

Метрики пилота

Матрица ролей

Числовые ориентиры — средние по пяти внедрениям RAG-поиска по корпоративным архивам в производстве и промышленных холдингах. Проекты закрыты соглашениями о конфиденциальности, данные приводятся обезличенно и не описывают ни одно внедрение по отдельности. Это наблюдения одного интегратора на небольшой выборке, а не отраслевая статистика.



Как пользоваться

СОДЕРЖАНИЕ

- | | |
|------------------------------------|------------------------------------|
| 1. Как пользоваться | 7. Метрики пилота |
| 2. Форма инвентаризации | 8. Матрица ролей |
| 3. Выборочная проверка сканов | 9. Что будет через полгода |
| 4. Чек-лист признаков неготовности | 10. Итоговый лист |
| 5. Протокол замера точки отсчёта | 11. Если своей команды под это нет |
| 6. Эталонный набор вопросов | |

Тетрадь заполняется по одному источнику за раз. Корпоративный архив почти никогда не один: СЭД, файловые шары подразделений, почтовые вложения, отдельные хранилища проектной документации. У каждого свои права, свои версии и своё качество, поэтому форму инвентаризации распечатайте в нужном количестве экземпляров.

Порядок работы такой. Сначала инвентаризация по каждому источнику — это даёт объёмы и форматы. Затем выборочная проверка сканов на тех источниках, где сканы вообще есть. Дальше чек-лист признаков неготовности, замер точки отсчёта и сбор эталонных вопросов. Роли назначаются в любой момент, но до начала пилота.

Заполняет тетрадь ИТ или владелец документооборота. Два раздела нельзя закрыть в одиночку: права доступа требуют участия ИБ, эталонные вопросы собираются только у пользователей, которые ищут документы каждый день.

Если данных по какому-то полю нет — так и напишите. Пустое поле с пометкой «не измеряли» честнее оценки на глаз и лучше видно на итоговом листе. Большинство незакрытых полей не мешают начать: они показывают, где неопределённость, и именно там позже возникают сюрпризы по срокам.

Три поля — исключение: схема прав доступа, владелец документов и точка отсчёта. Без них пилот запускать нечем, и на итоговом листе они вынесены отдельно.

Числовые ориентиры даны только в таблице метрик пилота. Это средние по пяти внедрениям RAG-поиска в производстве и промышленных холдингах, они нужны для сравнения порядка величин; ваши значения могут отличаться в обе стороны.

Форма инвентаризации

Источник: _____

Кто заполнял: _____ Дата: _____

Параметр	Как измерить	Значение
Общий объём	Число файлов и суммарный размер	
Доля сканов	Выборка 200–300 файлов, проверка наличия текстового слоя	
Качество текстового слоя	Ручная проверка выборки сканов: печати, рукописные пометки, перекос	
Глубина дублирования	Хеши содержимого по нормализованному тексту	
Наличие метаданных	Доля документов с датой, автором, подразделением, типом	
Форматы	Распределение по расширениям, включая архивы и вложенные архивы	
Права доступа	Как устроены сейчас: на уровне папок, ролей, отдельных документов	
Скорость прироста	Сколько документов добавляется в месяц	

Комментарии по источнику

Два поля, которые чаще всего оставляют пустыми. Глубина дублирования — потому что для неё нужно посчитать хеши, а не посмотреть на папку. Скорость прироста — потому что кажется несущественной. Первое определяет, сколько устаревших редакций попадёт в выдачу. Второе определяет, как часто придётся доиндексировать корпус, и это уже вопрос эксплуатации, а не запуска.



Выборочная проверка сканов

Раздел заполняется по тем источникам, где сканы есть. Цель — понять, какая часть отсканированных документов придёт в индекс с ненадёжным текстом.

Как отобрать выборку. Возьмите 200–300 файлов из источника случайным образом, а не подряд из одной папки. Документы из одной папки обычно сканировались одним человеком на одном аппарате в один период, и выборка получится нерепрезентативной. Если в источнике лежат документы разных лет — распределите выборку по годам пропорционально.

Что считать ненадёжным текстовым слоем. Признаки, по которым документ попадает в эту категорию:

Признак	Сколько документов выборки	Комментарий
Текстовый слой отсутствует полностью		
Печать или штамп перекрывает реквизиты		
Рукописные пометки поверх машинописного текста		
Перекося при сканировании более нескольких градусов		
Низкое разрешение, символы сливаются		
Многоколоночная вёрстка распознана в один поток		
Таблицы распознаны как несвязный текст		

Документ, попавший сразу в несколько категорий, считайте один раз — иначе сумма превысит объём выборки.

Как оформить результат. Посчитайте, сколько документов выборки попали хотя бы в одну категорию, и переведите в долю.

Размер выборки: _____ шт. Из них с ненадёжным текстом: _____ шт.

Доля от числа сканов: _____ % × Доля сканов в корпусе: _____ % = Доля от всего корпуса: _____ %

Ориентир: в среднем по пяти внедрениям ненадёжный текстовый слой имели более 20% сканов. Доля считается от числа отсканированных документов, не от всего архива, поэтому для итогового листа нужно именно правое значение.

Что с этим делать. Полностью автоматически задача не решается. Часть документов останется в статусе «найденно, но текст ненадёжен» даже после повторного распознавания с другим профилем. Заложите в проект отдельный этап на проблемные типы и решите заранее, будет ли такой статус виден пользователю. Видимый статус здесь лучше молчания: человек, получивший обрывочный текст без предупреждения, считает виноватой систему.



Чек-лист признаков неготовности

Отметьте по каждому признаку своё положение дел и то, что собираетесь с ним делать.

Сканы без надёжного текстового слоя

Наш случай: _____

Что делаем: _____

Документы без разметки

Всё, что оцифровывалось до появления регламента именования, приходит без даты, автора, подразделения и статуса. Отфильтровать выдачу по правам или отделить действующую редакцию от утратившей силу на таком материале нечем.

Наш случай: _____

Что делаем: _____

Параллельные версии одного документа

`_final`, `_final2`, `_испр`, `_испр_окончательный` в одной папке с разницей в полгода. Без дедупликации по нормализованному тексту и явной маркировки действующей редакции поиск найдёт старую версию с той же вероятностью, что и актуальную.

Наш случай: _____

Что делаем: _____

Разрыв между языком документов и языком запросов

Документ называется «Регламент взаимодействия при согласовании договорных обязательств», сотрудник спрашивает «как подписать контракт с новым поставщиком». На демонстрациях этого не видно, потому что демонстрируют на удобных вопросах.

Наш случай: _____

Что делаем: _____

Неописанные права доступа

Этот признак единственный, который не решается технически. Формально права есть — в виде наследования от папок и списков, накопленных за годы. Фактически на вопрос «кто из сотрудников что должен видеть» внятного ответа обычно нет ни у ИТ, ни у владельцев документов. Пока решение не принято на стороне заказчика, разграничение доступа в системе строить не из чего, и никакой подрядчик этого за вас не решит.

Кто принимает решение: _____

Срок: _____

Наш случай: _____



Протокол замера точки отсчёта

Без этого замера результат пилота нечем подтвердить: разговор сведётся к «мне кажется, стало лучше». Восстановить точку отсчёта задним числом невозможно.

Замеряются две величины: медианное время поиска документа и доля запросов, которые заканчиваются обращением к живому человеку. Вторая переводится в человеко-часы напрямую и лучше работает в разговоре с финансовым блоком.

Вариант первый: по журналам

Предпочтительный. Источники данных — журналы СЭД, заявки в архив, тикеты в службу поддержки.

Показатель	Значение	Период и источник
Медианное время поиска документа		
Максимальное время поиска		
Число обращений в архив за период		
Доля запросов с обращением к человеку		

Вариант второй: по наблюдению

Если журналов нет. Понаблюдайте за двумя-тремя сотрудниками в течение обычной рабочей недели и записывайте фактический маршрут поиска. Одна строка — один эпизод; за неделю их наберётся больше, чем строк в таблице, поэтому распечатайте лист в нескольких экземплярах.

Сотрудник	Что искал	Шаги маршрута	Затрачено времени	Нашёл

Предупреждение про опросы. Baseline, собранный анкетой, систематически завышен: люди помнят тяжёлые случаи и не помнят десяток быстрых. Если пользуетесь опросом — считайте полученное верхней границей и укажите это в отчёте, иначе результат пилота будет выглядеть лучше, чем он есть.

Эпизоды работы по устаревшим документам

То, что в человеко-часах не измеряется, но обычно и двигает проект. Вспомните случаи, когда подразделение отработало по отменённой редакции регламента или согласование ушло по заменённому шаблону договора.

В отчётности такой ущерб не виден, но в защите проекта он работает сильнее сэкономленных часов.



Эталонный набор вопросов

Самая дешёвая часть подготовки и самая пропускаемая. Набор нужен, чтобы измерить качество выдачи объективно, а не по впечатлению от демонстрации.

Правила составления. Вопросы формулируют пользователи, не проектная команда: команда невольно пишет так, как удобно системе. Набор заполняется до того, как кто-либо увидел первую выдачу, иначе он подстроится под наблюдаемое поведение. Несколько десятков вопросов достаточно для пилота.

Обязательно включите неудобные случаи: документы с похожими названиями, вопросы на границе двух регламентов и запросы к документам, которых в корпусе нет. Последнее нужно, чтобы проверить поведение системы при отсутствии ответа, — уверенный ответ без источника обходится дороже честного отказа.

Что писать в колонке «Тип»: обычный запрос, похожие названия, граница двух документов, запрос без ответа в корпусе, запрос со специфической терминологией.

№	Формулировка вопроса	Ожидаемый документ	Кто задал	Тип
1				
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				
13				
14				
15				
16				
17				
18				
19				
20				

№	Формулировка вопроса	Ожидаемый документ	Кто задал	Тип
21				
22				
23				
24				
25				
26				
27				
28				
29				
30				
31				
32				
33				
34				
35				
36				
37				
38				
39				
40				

Собранный набор пригодится дважды: на приёмке пилота и через полгода, когда понадобится показать, что качество не деградировало.



Метрики пилота

Зафиксируйте эти метрики до старта — тогда результат будет с чем сравнить. Ориентиры в правой колонке рассчитаны как средние по пяти внедрениям и приведены для сопоставления порядка величин.

Метрика	Что показывает	Ориентир по пяти внедрениям	Наше значение
Покрытие корпуса после загрузки	Доля документов, дошедших до индекса	73%	
Релевантность Тор-3	Доля вопросов набора с нужным документом в первых трёх	70%	
Релевантность Тор-5	То же для первых пяти	85%	
Доля ответов со ссылкой на источник	Проверяемость выдачи		
Медианное время поиска документа	Сравнение с точкой отсчёта	20% от прежнего	
Скорость индексации	Планирование первичной загрузки		

Как читать ориентиры. Покрытие важнее релевантности: документы, не дошедшие до индекса, пользователь не увидит и о них не сообщит. Релевантность считается по тому, что до индекса дошло, поэтому две метрики перемножаются. При покрытии 73% и Тор-3 в 70% от исходного корпуса в первых трёх позициях оказывается около половины запрошенного, если эталонный набор составлялся по документам, заведомо присутствующим в индексе. Проверьте, как собирался ваш набор, прежде чем цифра уйдёт в отчёт руководству.

Разрыв между Тор-3 и Тор-5 говорит больше, чем каждая метрика по отдельности. В наших разборах он составил пятнадцать пунктов: каждый седьмой правильный документ система находила, но ставила на четвёртую или пятую позицию. Это дефект ранжирования, а не поиска: нужный фрагмент система достала, поднять наверх не смогла. Лечится переранжированием, замена модели здесь ничего не даст. Практическое следствие: не обрезайте выдачу тремя результатами.

База сравнения

Отдельно зафиксируйте, с чем сравниваете. Ускорение поиска в пять раз, полученное поверх работающей системы электронного архива, и то же число относительно поиска по бумажным папкам — разные достижения. В наших разборах базой были уже работавшие электронные системы хранения.

С чем сравниваем: _____



Матрица ролей

Оценка упирается в людей быстрее, чем в технику. Заполните до начала пилота. В последней колонке — дата, когда человек подтвердил, что берёт роль; устная договорённость без даты на практике не удерживается.

Роль	Кто	Подразделение	Дата согласия
Владелец документов — знает, какая редакция действующая, и имеет право это решать			
Представитель ИБ — согласует контур и схему разграничения доступа			
Пилотная группа пользователей — ищут документы ежедневно, готовы еженедельно смотреть на выдачу			
Владелец корпуса на эксплуатацию — подключает источники, следит за доиндексацией, разбирает жалобы			

Про третью роль. Проекты, где пилотную группу не собрали, доходят до демонстрации и останавливаются. Демонстрация нравится всем, а обратную связь по реальным запросам давать некому.

Про четвёртую роль. Она понадобится позже, но назначать её нужно на старте: за полгода эксплуатации корпус вырастает настолько, что целиком его уже никто не просматривал, и у него должен быть хозяин.



Что будет через полгода

Цифры этого раздела нужны для бизнес-кейса. Все они — средние по пяти внедрениям, замер через шесть месяцев после запуска.

Активными остаются не все. Через полгода систему регулярно использовали 34% целевой группы — тех, кому выдали доступ и кого учили. Ожидание, что искать через новую систему будут все, не сбывается: часть сотрудников обращается к документам раз в квартал, часть продолжает пользоваться привычным маршрутом, пока он работает. Считайте экономию от этой доли, а не от списочной численности подразделения.

Обращений становится больше. Их число за полугодие выросло на 62%. Люди начали искать чаще, потому что искать стало дешевле: запросы, от которых раньше отказывались, вернулись в работу.

Отсюда арифметика, которую лучше проговорить с финансовым блоком заранее. Один поиск занимает 20% прежнего времени — экономия 80%. Но суммарное время подразделения сокращается примерно до трети от прежнего, потому что обращений стало в полтора раза больше. Разрыв между 80% и примерно 68% — подавленный спрос, которого до внедрения не было видно в отчётности. Покажите его сами, пока это не сделал финансовый директор.

Корпус растёт быстрее ожиданий. Объём индекса за полгода вырос с 40 тысяч документов до 600 тысяч. Это и хорошая новость — проект дошёл от пилотной выборки до реального архива, — и повод заранее назначить владельца корпуса.



Итоговый лист

Сведите заполненное и определите свой случай.

Что должно быть собрано	Готово
Форма инвентаризации заполнена по каждому источнику	
Выборочная проверка сканов проведена, доля от корпуса посчитана	
Чек-лист признаков неготовности заполнен	
Схема прав доступа: есть ответ «кто что должен видеть»	
Владелец документов назначен поимённо	
Точка отсчёта зафиксирована	
Эталонный набор вопросов собран	
Остальные роли назначены и согласованы	

Три строки, выделенные жирным, блокирующие: без них пилот запускать нечем.

Пилот можно начинать

Все восемь строк отмечены. Признаки неготовности либо отсутствуют, либо по каждому записано, что вы с ним делаете. Владелец документов и представитель ИБ назначены поимённо.

В этом случае разговор с подрядчиком начинается сразу с объёма работ и сроков, а заполненная тетрадь служит техническим заданием на оценку.

Начинать можно с оговорками

Инвентаризация и точка отсчёта закрыты, но один или два признака неготовности отмечены как проблемные и плана по ним пока нет. Чаще всего это сканы и метаданные.

Такой пилот запускается, но берите узкий срез: один-два типа документов, одно подразделение, понятный набор пользователей. Проблемные признаки заведите в план работ отдельными строками со сроками, иначе они всплывут на третьем месяце.

Сначала нужно закрыть пункты

Не отмечена хотя бы одна из трёх выделенных строк: права доступа не описаны и решение принимать некому; не назначен владелец документов, способный сказать, какая редакция действующая; точка отсчёта не зафиксирована.

Первое — организационная задача, техникой не решается. Второе определяет, сможете ли вы вообще отличить актуальный документ от отменённого. Третье означает, что результат пилота будет нечем измерить, и разговор о его продлении превратится во вкусовщину.

Наш случай: _____

Что закрываем и к какому сроку: _____



Если своей команды под это нет

Оценка корпуса не требует подрядчика: всё описанное здесь выполняется силами ИТ и владельцев документооборота. Мы выложили методику целиком именно поэтому.

Но если людей под это не выделить или хочется пройти оценку с теми, кто делал её раньше, — AZONE-AI может провести её вместе с вашей командой. Инвентаризация источников, выборочная проверка качества сканов, разбор схемы прав доступа и замер точки отсчёта. На выходе — заполненные формы, эталонный набор вопросов и архитектура пилота в закрытом контуре с реалистичными сроками.

Развёртывание идёт на инфраструктуре заказчика: модель, векторное хранилище и документы не покидают периметр. Для аттестованных сегментов это условие обязательное, и обсуждать его лучше на этапе оценки, а не при защите архитектуры.

Обсудить оценку архива

Подробный разбор методики: статья «Как оценить корпоративный архив перед RAG-проектом» на сайте AZONE-AI

azoneai.ru/contact · услуга «Внедрение RAG» · лицензии ФСТЭК, ФСБ, МО РФ · статья: azoneai.ru/blog/32-ocenka-arhiva-pered-rag

AZONE-AI · направление компании АЗОН Корпоративные AI/LLM-решения в закрытых контурах: on-premise развёртывание, RAG-системы, ИИ для информационной безопасности

Числовые ориентиры в документе — средние по пяти внедрениям RAG-поиска по корпоративным архивам в производстве и промышленных холдингах. Проекты закрыты соглашениями о конфиденциальности, поэтому данные приводятся обезличенно.