



# КОРПОРАТИВНЫЙ ПОИСК – единая точка доступа к знаниям Компании



# Корпоративный поиск



Объединяет источники данных, информационные системы и каталоги, аналитические отчеты, хранилища структурированной и неструктурированной информации, становясь **единым окном доступа к данным**.

Развитие поисковой платформы дает толчок к развитию направления **семантического анализа** в Компании



## Легенда

- ◆ Приложения по анализу контента, реализованные на базе платформы корпоративного поиска
- ◆ Информационные системы

Инструменты семантического анализа, предоставляемые платформой поиска, позволяют **не только находить информацию, но и выявлять скрытые связи и знания**

Как мы привыкли жить?

>180

порталов и ИС

>350Т6

в файловых папках

>116 650 000

файлов в ИС ГПН

В каком документе содержится нужная информация?

Где необходимо искать требуемый документ?

Какая версия документа актуальная?

Как правильно сформулировать запрос?

В мире Корпоративного поиска:

## ЕДИНОЕ ОКНО И УНИВЕРСАЛЬНЫЙ ИНСТРУМЕНТ ПОИСКА

- Интеллектуальный поиск контента, данных, сервисов из корп. систем.
- «Поиск как сервис» - встраивание умного поиска в ИС компании.

## ПОИСК ДОСТУПЕН ВСЕМ СОТРУДНИКАМ КОМПАНИИ

- Доступ с порталов подразделений

Корпоративный поиск

Портал  Сотрудники  Корп. поиск

- Доступ по прямому адресу

## НОВЫЕ ВОЗМОЖНОСТИ ПРИ ВЫПОЛНЕНИИ ПРИВЫЧНЫХ ЗАДАЧ

- Поиск по периодике, новостям, официальной документации.
- Поиск по корпоративным сервисам.
- Изучение предметной области – поиск близких по тематике документов.
- Улучшенный поиск по сотрудникам.
- Поиск по аналитическим отчетам.

• **Сохранение накопленных знаний** в Компании.

• **Исключение двойного финансирования** при запуске новых проектов.

• **Поиск с учетом корпоративного языка.**

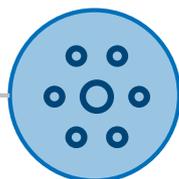
• **Сокращение времени на поиск** необходимой информации.

• **Агрегация лучших практик.**

• **Упрощение доступа к данным** за счет интеграции с инструментами функции Управления данными



Целевой охват аудитории – **70 тыс.** сотрудников!



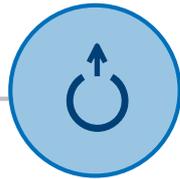
**4,5 из 5** – оценка пользователей уровня доверия к Поиску;  
**4,6 из 5** – удобство использования;  
**4,4 из 5** – релевантность выдачи.



Количество документов в системе приближается к 5 миллионам: пользователям доступны **4 780 244** объектов.



Корпоративный поиск работает в продуктиве всего **14** месяцев, за которые команда успела выпустить **10** релизов, постоянно развивая функциональность инструмента.



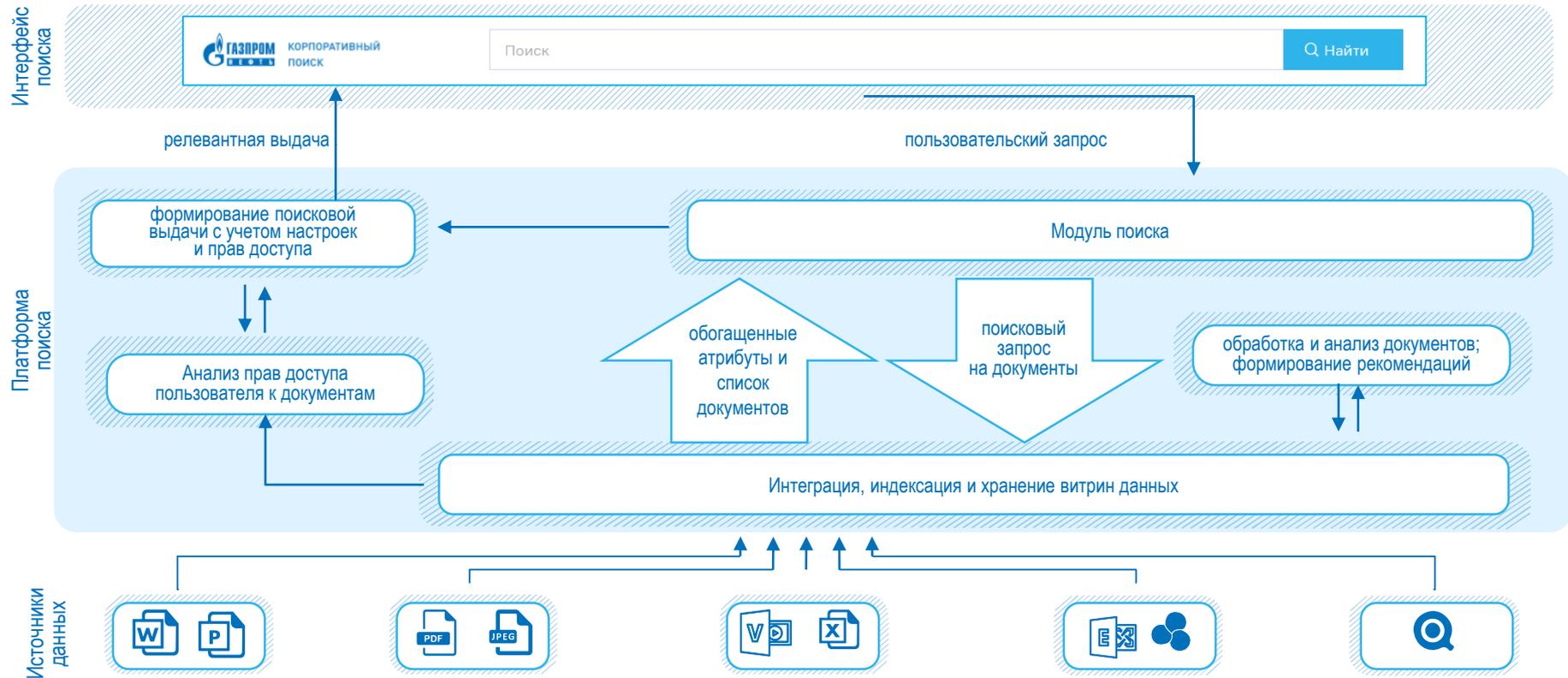
Корпоративный поиск регулярно индексирует **27** систем-источников. Среди них – **5** внешних сайта, контент которых собирает специально разработанный краулер.



Задано более **411 тыс.** поисковых запросов



Над продуктом ежедневно работают **13** членов внутренней команды разработки





индексация источника Корпоративным поиском



встраивание поиска в поисковую строку на вашем портале



аналитические инструменты (методы)

- Индексация данных источника полностью или выбранных разделов.
- Источник появляется в выдаче Корпоративного поиска.
- Пользователи, найдя информацию в Корпоративном поиске, переходят непосредственно в источник, что повышает конверсию и информированность пользователей о полезных материалах источника.

- Индексация данных источника полностью или выбранных разделов.
- Встраивание переадресации поискового запроса из системы в Корпоративный поиск.
- Поиск возвращает релевантную выдачу по источнику, с фильтрацией, автоподсказками, подсветкой в результатах обработки запроса.

- Метод возвращает ключевые слова и аннотацию по документу/объекту источника.
- Метод формирует к документу источника релевантный семантически близкий набор документов.
- Метод возвращает наборы релевантных запросу по ключевым словам документов, с ссылками на источники документов, базовыми параметрами документов.

 <p>стандартные коннекторы</p>	 <p>веб-сервисы</p>	 <p>вспомогательные сервисы</p>	 <p>краулеры</p>
<ul style="list-style-type: none"><li>• Использование стандартных коннекторов корпоративных систем, которые идут к ним «из коробки».</li><li>• Актуально для enterprise систем типа SharePoint, IBM Lotus, IBM FileNet.</li><li>• Любой описанный интерфейс/ресурс с разработанным ранее API.</li></ul>	<ul style="list-style-type: none"><li>• Совместная разработка веб-сервисов для передачи и приема данных от системы источника в Корпоративный поиск.</li><li>• Целесообразно для очень кастомной интеграции, наличия сложных полей, а не просто индексации текстовых страниц (см. вариант с краулерами).</li></ul>	<ul style="list-style-type: none"><li>• Используется КШД или DFS, как временное хранилище для передачи информации.</li><li>• Целесообразно, когда нет возможности сделать прямую интеграцию указанными выше методами и возможна интеграция через файлы с выгрузками в заданном формате или копии витрин данных.</li></ul>	<ul style="list-style-type: none"><li>• Использование краулеров для сбора данных с порталов и систем, использование Apache Nutch или самописных «пауков» по специфике источников.</li><li>• Актуально для внешних систем и типовой индексации страниц порталов без коннекторов.</li></ul>

# Pipeline обработки файла



## 1. Получение файла

Копируем файл из источника, сохраняется во временной директории на время извлечения контента.

1

## 2. Преданализ

Анализ типа файла: формат, расширение, архив или файл для выбора подхода по обработке (contentType, mimeType)

2

## 3. Извлечение контента

Разархивация, Tika, OCR, структуризация для спец форматов. Превращаем в извлечённый текстовый образ документа.

3

## 4. Обогащение

Извлечение метаданных из файла (автор, дата создания, изменения), дополнительные параметры из пути расположения и тд.

4

## 5. Сохранение

Извлеченный текстовый образ файла и метаданные записываются в MongoDB. Скопированный файл удаляется из временной директории.

5

## 6. Токенизация

Текстовый образ файла проходит морфологический и синтаксический разбор в вычислительном кластере.

6

## 7. Векторизация

В вычислительном кластере происходит построение эмбединга (семантического вектора) для текстового образа файла

7

## 8. TextMining

Выделение ключевых слов, аннотаций из полученного документа.

8

## 9. Каталогизация

На основе предобученной модели ML происходит классификация текстов на каталог данных.

9

## 10. Семантический поиск

Расчет cosine-similarity на пространстве семантических векторов с применением алгоритмов LSH для эффективного расчета смысловых пересечений на коллекциях из миллионов документов.

10

## 11. Индексация

На основе обогащенного текстового образа документов в ElasticSearch строится полнотекстовый индекс для поиска и фильтрации документов.

11

# Как работает Корпоративный поиск?



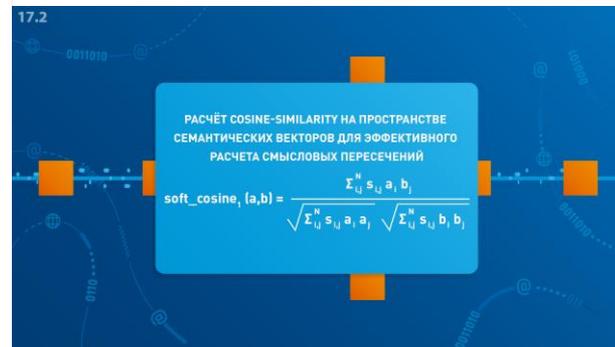
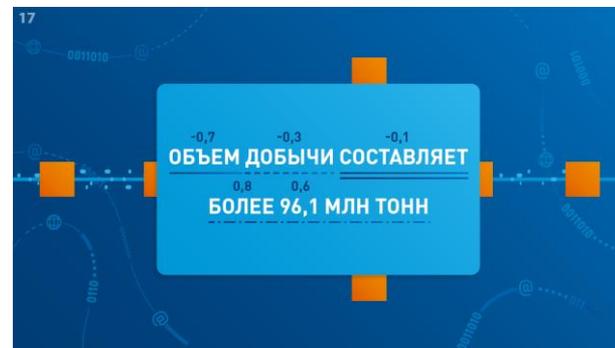
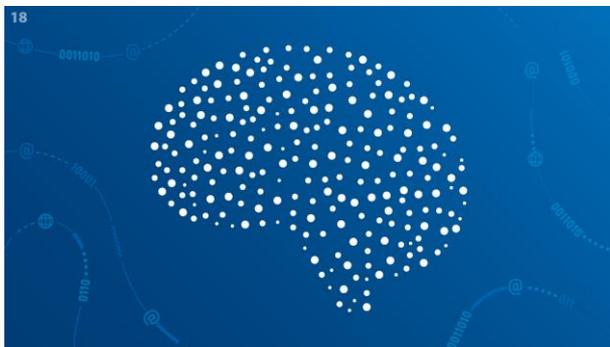
СТРЕМИМЯ  
К БОЛЬШЕ

Корпоративный  
поиск

SEARCH GAZPROM-NEFT.LOCAL

Находить информацию проще!

Нажмите для просмотра



## Команда активно развивает продукт:



Подключая новые источники контента



Разрабатывая новый функционал



Создавая удобные продукты-спутники и сервисы

Ваши голоса **очень важны** для нас!