НЭБ

Национальная электронная библиотека Российской Федерации

R&D - 2020

Публичный **отчет о работе** подразделений Российской государственной библиотеки “**Лаборатория исследований и разработки**” и “**Отдел инженеров знаний**”за **2020** год

Стратегия начального этапа работы R&D

Работа R&D подразделения Российской государственной библиотеки началась конце 2019 года.

В качестве основных стратегических принципов, упрощающих моменты выбора направления работ и используемых решений:

1. **Увеличение доступности цифровых объектов**. Речь идет о доступности в самом широком смысле - технической, юридической, смысловой, возможность найти нужную публикацию и т.п.
2. Максимальную автоматизацию рутинных операций при работе с информацией и материальными фондами, по сценарию, когда **автоматика выполняет ассистирующую человеку роль**,
3. **Базовой** сущностью проектирования и анализа задач являются модели **жизненного цикла данных**.
4. Работа с **большими объемами данных** должна начинаться с применения решений, дающих “**обзор с птичьего полета**”, они позволяют избежать неожиданных вариаций и дают понимание, на чем фокусироваться при работе с ними.
5. Ввиду очень высокой фрагментированности экосистемы данных, связанных с издательским и библиотечным делом, вторым основным принципом стало **обеспечение максимальной совместимости решений с существующими стандартами** цифровых публикаций, метаданных, каталогов, обмена данными, защиты информации.

   Источниками подобных стандартов являются ГОСТ, ISO, W3C, IFLA, IEC, US LC и прочих международных авторитетных организаций.

   Принятие международных стандартов и их совместная доработка и корректная адаптация - очень сложный, но единственный масштабируемый путь, способствующий обмену высоко систематизированным опытом, делающим эффективными как централизованное управление и планирование, так и децентрализованную координацию любой конструктивной и коммуникативной деятельности.

Проект системы дистрибуции цифровых изданий

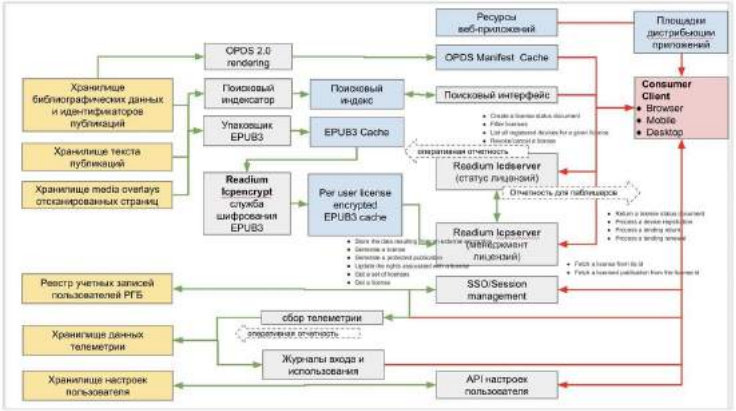
Одной из начальных задач стала разработка архитектуры и прототипа цифровой системы дистрибьюции электронных изданий на базе СПО с не привязанной к конкретному поставщику DRM системой.

За основу была взята платформа EU Digital Reading Lab - Readium, были проработаны схемы интеграций с поставщикмии программными клиентами. В качестве основного формата дистрибьюции и хранения был выбран EPUB3.

В рамках апробации развернута референсная платформа Readium, включающая сервер лицензий, нестандартный механизм шифрования, стример и ряд других служб. Оценочное количество юридических и операционных проблем после начала массовой эксплуатации системы бытро увеличивалось.С целью решения части из них были изучены и пробировнны на предполагаемых типовыми лицензиях модели автоматизированного управления цифровыми лицензиями (подмножество ODRL2)

Результаты:

- Разработана базовая архитектура ансамбля, описаны точки интеграций, разработано несколько архитектур подсистем тактического уровня, которые в дальнейшем были использована при разработке Системы Интеграции Данных
- Было принято решение о заморозке проект и апробации ряда подсистем в составе проекта “НЭБ Свет” в результате детализации SWOT факторов проекта.
- Вступление оператора НЭБ в европейский консорциум Readium



Цифровой пре-паблишинг и ре-паблишинг

Кризис четвертого сектора экономики в 2020 году серьезно затронул издательства. Особенно экономически пострадали малые и специализированные (non-fiction) издатели, которые не смогли перейти на современные технологии производства и распространения цифровых изданий.

При анализе поставленных различными дистрибьюторами образцов EPUB, а также текущих распространенных техпроцессов было выявлено крайне низкое качество сборки контейнера публикаций, верстки, метаданных. Массово использовался формат PDF, ориентированный на печать с постраничной и часто многоколоночной верстки, это делает неудобным чтение с экрана, особенно в случае мобильных платформ.

Начальным видением решения стала концепция цифрового **цифрового ре-паблишинга**, когда из произвольных исходных цифровых форматов собираются единообразные для конкретной платформы и высококачественные цифровые публикации.Для этого система осуществляет структурный и визуальный анализ, выявляет и сохраняет закономерности в разметке.

Решение относится к направлению нашей R&D активности в облсти т.н. **Semantic Publishing**, когда управление содержимым публикации полностью независимо от управления ее конкретным представлением.

“Выделить целевой грант на создание типового сайта интернет-магазина для независимых книжных магазинов, интегрированной национальной общей системы интернет-торговли и продвижения. Это поможет маленьким книжным пройти цифровую трансформацию и подключиться к онлайн-торговле и логистике”

Цитата из публично размещенной петиции https://www.change.org/p/петиция-представителей-книжного-бизнеса-россии?use_react=false

Цифровой пре-паблишинг и ре-паблишинг

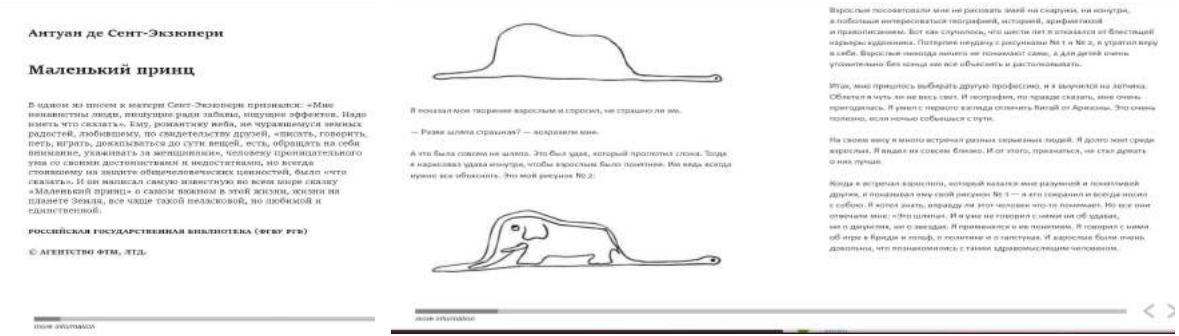
Система задействует спектр логик по пересборке спектра входных форматов, включая склейку PDF и прочих неадаптивных постраничных представлений.

Дополнительный набор эвристик уровня навигации формирует многоуровневые оглавления и пересобирает концевые сноски.

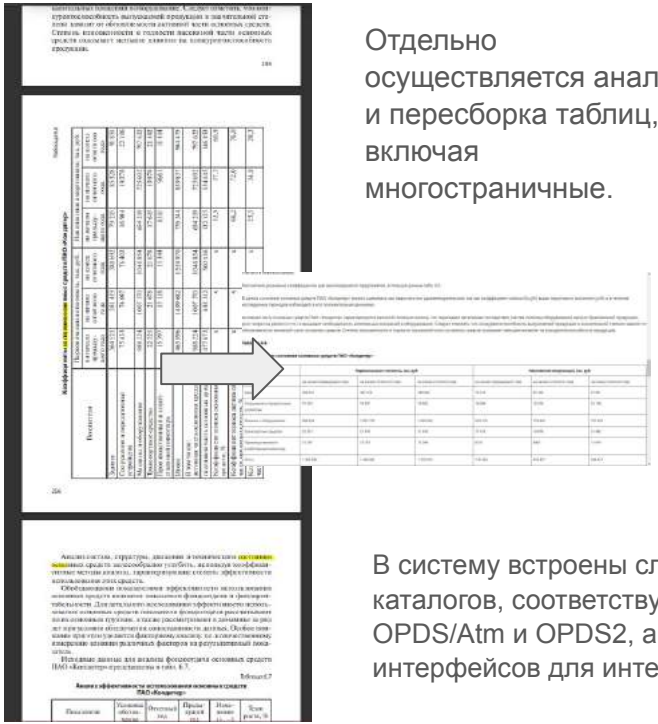
Ко внешним ссылкам применяются дополнительные механизмы санации.

Автоформатирование текста (подсистема санации) старается сохранить структуру стилей и акцентных выделений в тексте, при этом конвертируя “мимикрирующую верстку”, когда, например, параграфы разделяются двойным символом новой линии или центрирование делается пробелами на корректную HTML структуру.

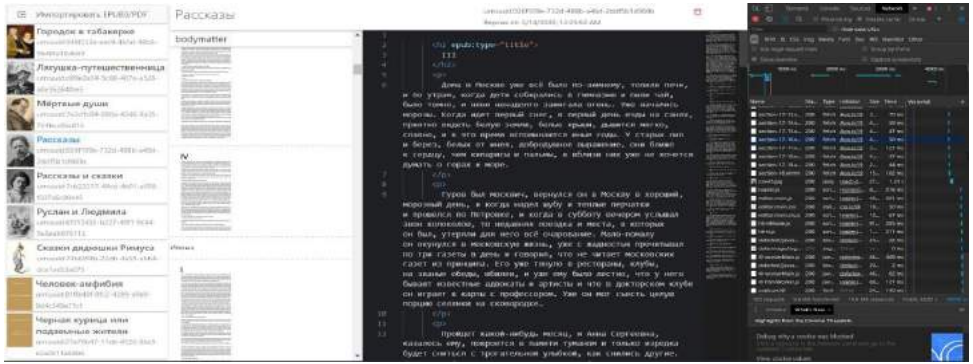
Переформатирование текстов гибко конфигурируется, базовый режим в своей работе ориентирован на лучшие практики и образцы “веб-типографики”, а также руководства по визуальному стилю приложений для мобильных платформ. Значительную часть работы берет на себя высококачественная программная библиотека “Типограф” (https://www.npmjs.com/package/typograf, автор - Денис Селезнев).



Цифровой пре-паблишинг и ре-паблишинг



Отдельно осуществляется анализ и пересборка таблиц, включая многостраничные.



Интерфейсная подсистема компактного рендеринга (схожая с просмотром результата автоконвертации в Kindle Create) упрощает навигацию по всему объему публикации. Имеется режим рендеринга вида оригинальным механизмом клиентской части платформы Readium (1 поколения)

В систему встроены службы web каталогов, соответствующих стандартам OPDS/Atm и OPDS2, а также ряд других интерфейсов для интеграции.

Цифровой пре-паблишинг и ре-паблишинг - результаты

- Система использована при работие над приложением “НЭБ Свет” https://svetapp.rusneb.ru/
- Система продемонстрирован основным игрокам рынка цифрового издательства РФ и получила крайне положительные отзывы.
- Публичная демо-версия системы ре-паблишинга для демонстрации процесса и результатов поставщикам https://pub-maker.rusneb
- Также реализован прототип интерфейса редакторского стола для сборки новых публикаций.

Программный код системы достаточно трудоемок для подготовки к публикации в открытый доступ, но это не исключено в случае массового интереса к системам подобного назначения или к конкретно нашей реализации.

Структурная модель цифрового издания

Для унификации данных, описывающих содержимое цифровой публикации был необходим Lingua franca, который был бы основным форматом хранения и связующим звеном между форматами авторинга и дистрибуции, а также оптимизированных под платформу представлений. К тому же высокая степень соответствия общемировым стандартам снижает расходы на документирование и интеграцию с системами.

Мы принимаем активное участие в работе W3C по формированию нового стандарта данных цифровых публикаций “Web publication manifest”, так как мы столкнулись с рядом серьезных проблем в рабочем варианте стандарта, препятствующих его использованию, Одной из них была высокая степень его фрагментированности (4 активно разрабатываемых несовместимых версии). Комитету W3C)была оказана помощь в анализе соответствия вариаций рабочего стандарта. Также мы предоставили детальный анализ около десятка логических или процессуального рода несоответствий в формате метаданных, варианты альтернативных решений.

Чтобы консолидировать эти стандарты и ряд семантических систем W3C для прикладного использования нами была сформирована **полная цифровая модель издания**, описывающая его от уровня тома до секций и ниже, вплоть до символов, предполагая использования на низком уровне разных реализаций **DOM моделей** (Document object model), в качестве основной модели дом была выбрана модель ProseMirror Content Model от R&D подразделения NY Times. Высокоуровневая структура опирается на неформализованный онтологический словарь W3C EPUB3 Structural semantic vocabulary. Итоговая JSON схема и упрощенные структурные схемы размещены в публичный доступ:

Результаты:

- Разработана и опубликована схема цифровой модели издания и доп инструменты и материалы https://github.com/ndlrf-rnd/digital-publicatuion-schema
- Разработаны конверторы из близких форматов: Readium manifest, EPUB3 manifest, FB2 metadata, W3C webpub manifest, W3C webpub. Разработаны максимально интероперабельные схемы валидации метаданных.
- Взаимодействие с W3C, результатом которого стало консолидация стандарта веб публикаций и коррекция комитетом ряда выявленных нами проблем. (ряд публично доступных эпизодов публично доступного взаимодействия: https://github.com/w3c/pub-manifest/issues/203, https://github.com/w3c/pub-manifest/issues/202, https://github.com/w3c/wpub/issues/465)
- Планируется формализация и контрибуция этого материала в тело стандартов W3C и платформы Readium.

Система Интеграции Данных (СИД/SEED)

$$LT = \{(y,x) \in Y \times X | (x,y) \in L\}$$

При решении проблемы поставки данных для проекта Реестр Книжных Пмятников (https://knpam.rusneb.ru/), а также его синхронизация с разделом НЭБ “Книжные Памятники” ()https://kp.rusneb.ru) для широкой аудитории возникла необходимость массовой унификации фрагментов каталогов РГБ (MARC 21) и РНБ (RUSMARC).и множество дальнейших преобразований к форматам данных, используемых/знакомым внешним подрядчикам, а также передачи сканов с метаданными в цифровой архив Министерства культуры.

Это был отправной сценарий работы СИД на начало 2020 года. С тех пор было апробировано множество вариантов внутреннего хранения данных, стабильно добавлялась поддержка новых форматов и стандартов.

Сейчас СИД все еще находится в экспериментальном режиме и частично доступен из публичной зоны Интернет по адресу: https://catalog.rusneb.ru и выполняет группы функций:

- Master data management / Data integration
- Автоконвертации различных технических и логических форматов метаданных (crosswalks ecosystem)
- Проксирование обращений к объектному хранилищу цифровых публикаций и прочих медиа-объектов, предоставляющая доступ к цифровым объектам в рамках единого API
- Предоставление рабочей зоны для связывания данных от разных поставщиков.
- Единого персистентного журнала для метаданных, поступающих в ряд проектов НЭБ данных
- Консолидированного электронного каталога OPDS2
- Планировщика технических операций по серийному импорту архивов.
- Аналитического инструмента

СИД — Поддерживаемые форматы данных

Форматы сериализации:

- ISO 2709:2008 (локальная адаптация: ГОСТ 7.14-98 (ИСО 2709-96) СИБИД. Формат для обмена информацией. Структура записи)
- MARCXML

Стандарты описания:

- MARC 21 Bibliographic
- MARC 21 Authority
- MARC 21 Holdings
- RUSMARC Bibliographic
- UNIMARC Bibliographic

ONIX for books:

- ONIX for books 3.0
- ONIX for books 2.x (ограниченная поддержка)

Стандарты описания оцифрованных изданий:

- METS 1.12.1
- Alto 4.2
- Page.xml (частичная поддержка)
- hOCR 1.2

Метаданные электронных изданий:

- EPUB 3.X
- FB2 (экспериментльня поддержк)
- W3C Publication Manifest

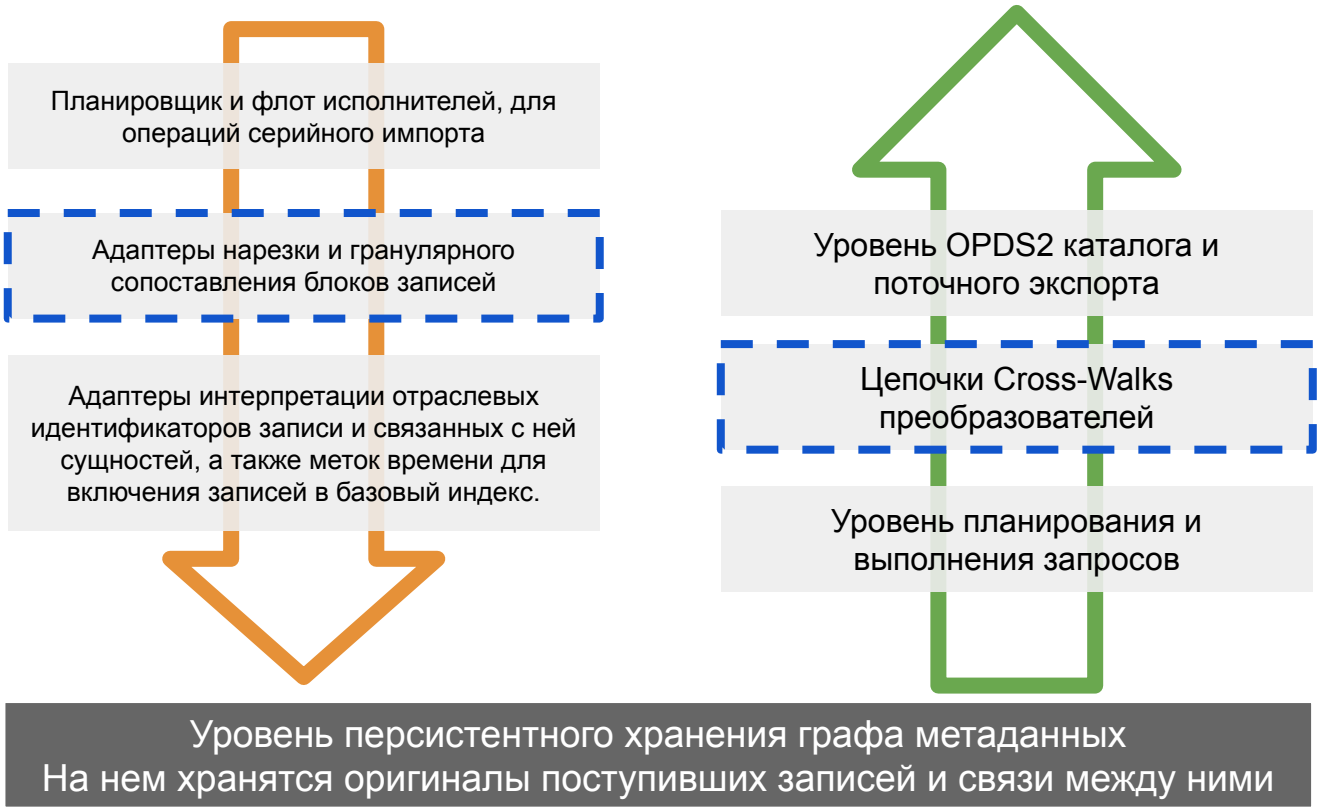
Проприетарные форматы:

- Yandex Market Language
- Платформа РАН ras.jes.su

Онтологии:

- SKOS
- BIBFRAME 2
- Schema.org
- WordNet (глоссарий WordNet Glossary)
- EPUB 3 Structural Semantics Vocabulary (В рамках структурной модели цифрового издания)

СИД — архитектура ядра



СИД — План публичного размещения кода системы

Реалистично оценив, что будет более полезно для специалистов: стабильные и покрытые тестами конверторы популярных форматов метаданных или слабо документированный стек логики общего назначения?

Мы сделали очевидный выбор по отделению от СИД подсистемы с адаптерами и ее публикацию в качестве начальных шагов.

У нас не получится быстро избавить логику конверторов от нестандартизированных условий и правил, несмотря на то, что при разработке мы старались в явном виде не оперировать с “домашними” кодами и расширениями полей данных.
Адаптеры больших и сложных форматов вроде MARC или ONIX также содержат формализованные схемы для JSON представления, машинночитаемую документ

	Closed beta / partners	Public Beta
Преобразователи форматов семейства MARC, включая RUSMARC -> MARC21	Q1 2020	Q1 2021
Конверторы форматов семейства ONIX	Q3 2020	Q1 2021
Внешний семантический индекс	Q3 2020	Q1 2021
Анализатор исходных полей MARC (ONIX?)	Q4 2020	Q1 2021
Веб-навигатор (И автоформы?)	Q1 2021	Q2 2021
Ядро СИД	Q4 2020	Q3–Q4 2021

СИД — Технические детали

Как целевая ERD модель предполагаются анлоги IBM Atomic Warehouse Model
https://www.ibm.com/support/knowledgecenter/en/SS9NBR_9.1.0/com.ibm.ima.using/comp/awm/intro.html

Используется расширенная мультитемпоральная модель времени схожая с Valid Time Model (https://en.wikipedia.org/wiki/Valid_time)

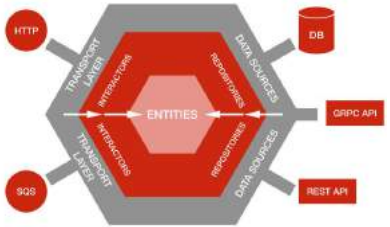
Формируемые структуры данных близки к аналитическому решению Chronos engine(https://slideplayer.com/slide/7100058/)

В качестве внутреннего языка запросов используется трансляция Cypher -> PSQL

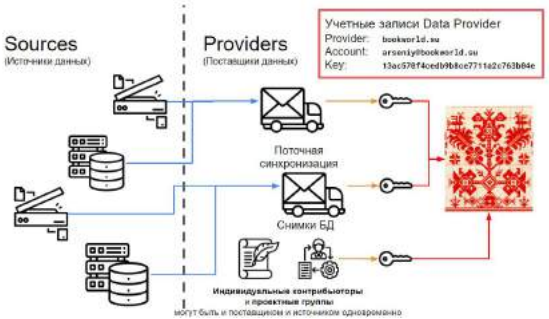
Слой персистентного хранения обеспечиватся оверлейным интерфейсом и транслятором запросов поверх СПО СУБД - PostgreSQL и RocksDB/LevelDB (выбор конкретного LSM Tree решения зависит от приоритетного сценария использования и статистики по количеству обновления ревизий отдельной записи)

Уже после первых апробаций СИД на реальных проектах вышла статья о “Гексагональной архитектуре”.от специалистов компании Netflix US.
Она ближе всего к верхнеуровневому описанию архитектуры системы.

Источник:
https://netflixtechblog.com/ready-for-changes-with-hexagonal-architecture-b315ec967749



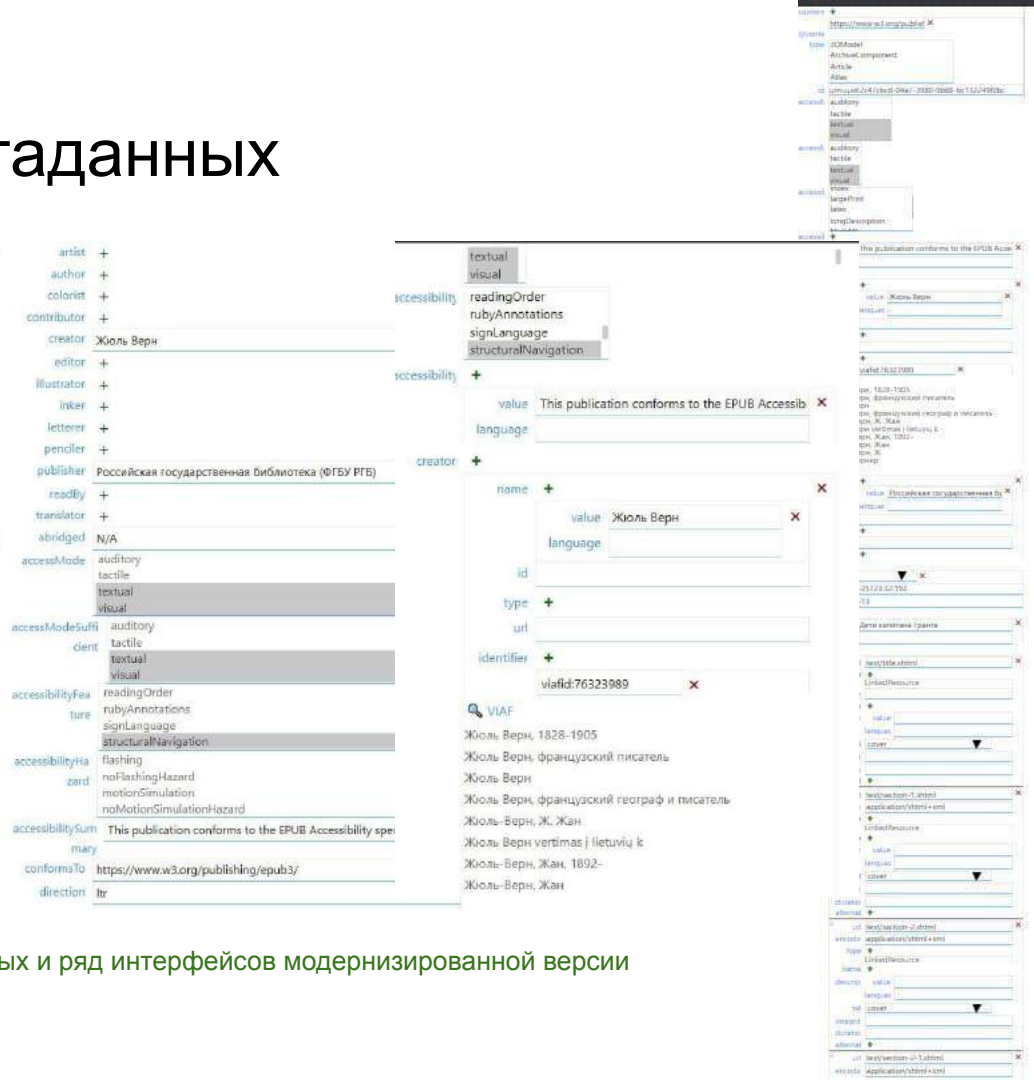
СИД явно разграничивает **технического поставщика данных** и **источник данных**



Визуальный редактор метаданных

В силу значительного количества различных форматов обмена данными часто требовался минимально рабочий WISIWYG редактор. Учитывая, что стандарты вроде MARC21 могут иметь комбинаторную сложность до десятка тысяч полей, имеющих отдельный смысл, наивные подходы работали плохо.

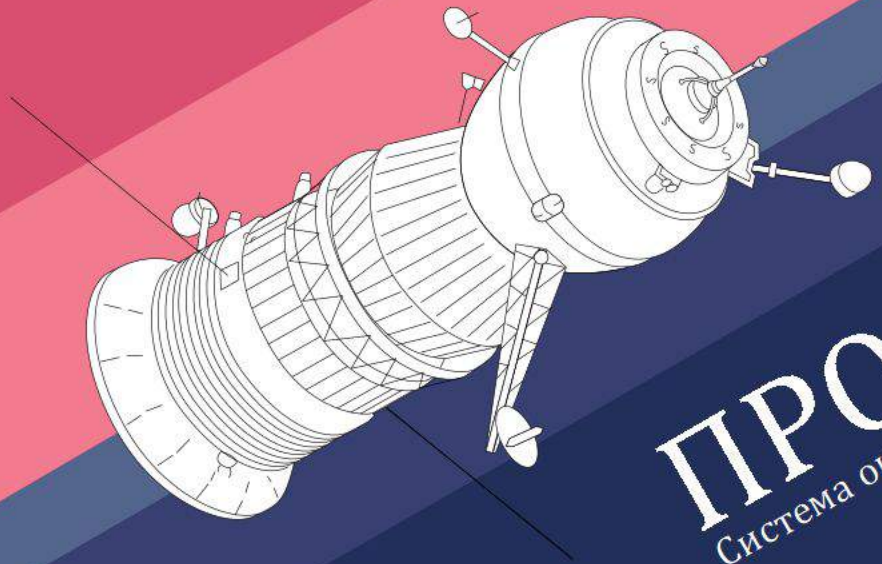
Был сделан прототип системы, автоматически формирующей интерфейс для редактирования любых данных на базе JSON schema. Большое количество эвристик почти полностью убирают “мусорные” опции и итоговое дерево форм схоже с интерфейсом, разработанным человеком. При разработке была изучена история развития система автогенерации форм Microsoft, IBM, Oracle, SAP. Концептуальная апробация осуществлялась на метаданных Web Publication/EPUB3 и MARC21



Результаты:

- Ограниченное внутреннее использование
- Планируется для внедрения в Систему интеграции данных и ряд интерфейсов модернизированной версии Реестра книжных памятников

Оцифровка периодических изданий



ПРОГРЕСС-1.0.0

Система оцифровки архивных документов

«Прогресс» в оцифровке изданий

Проект по оцифровке периодических изданий возник как инициатива руководителя проектного офис НЭБ, изначально сфокусированння на ведомственных архивах с целью осуществления массовой событийной реконструкции биографий людей.
Несмотря на отзывчивость фондодержателей, помимо очевидной проблемы распознавания рукописного ввода, стали очевидны множественные юридические проблемы, снижающие потенциал и доступность результата и программа была переориентирована на периодические издания.

Архитектурно «Прогресс» изначально задумывался, как модульная система, где отдельные модули представляют из себя автономные и самоценные CLI утилиты или службы. При этом в изначальное видение входила “turnkey” сборка системы, которая бы “из коробки” выполняла задачу по преобразованию папки с изображениями в папку с цифровыми изданиями.

«Прогресс» – Предобработка растра

- Коррекция яркости и контрастности, детекция разрешающей способности при захвате растра
- Нейросетевая детекция текста на изображении
- Коррекция ориентации, линейных искажений

«Прогресс» – Анализ структуры документа

- Нейросетевая детекция контентных и разделительных сегментов на странице. Формирование векторных границ и начальных гипотез об их типах.
- Индуктивный анализ, выдающий окончательное решение о расположении блоков.
- Эвристический анализ “Порядка следования” и “порядка чтения” блоков документа.

«Прогресс» – Распознавание текста

- Нарезка документа на серию изображений содержимого блоков.
- Экспорт пакета контент-блоков для внешних средств анализа.
- Осуществление OCR для текстовых блоков
- Реконструкция артефактов технологии halftone printing для изображений

«Прогресс» – Анализ текста

- Разрешение матрицы гипотез OCR средствами авторегрессионных языковых моделей или векторизованной имплементацией алгоритма класса BEAM search
- Эвристическая посткоррекция: удаление ложных символов, преобразование переносов, сегментация на параграфы.
- Разрешение гипотез о блоках заголовков и метаданных уровня статей
- Построение постатейного индекса сводного и по выпускам
- Выделение блоков атрибуции авторства и времени написания, примечаний и сносок (Экспериментальная технология)

«Прогресс» – Анализ сущностей (named entities linking)

- Построение первичного индекса имен сущностей (NER Index)
  - Формирование первичного индекса именованных сущностей средствами авторегрессионных или transformer моделей.
  - Экспорт первичного NER индекса в формате JSON
- Индекс связанных сущностей (LNER index)
  - Локализация и интерпретация темпоральных языковых конструкций.
  - Консолидация символов первичного индекса именованных сущностей, контекстуально зависимое совмещение и разделение отдельных символов, т.н. entity linking.
  - Экспорт LNER индекса в формате JSON.

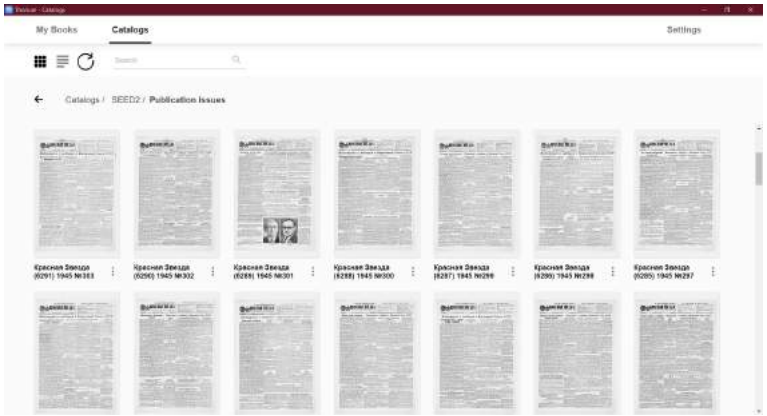
«Прогресс» – Экспорт

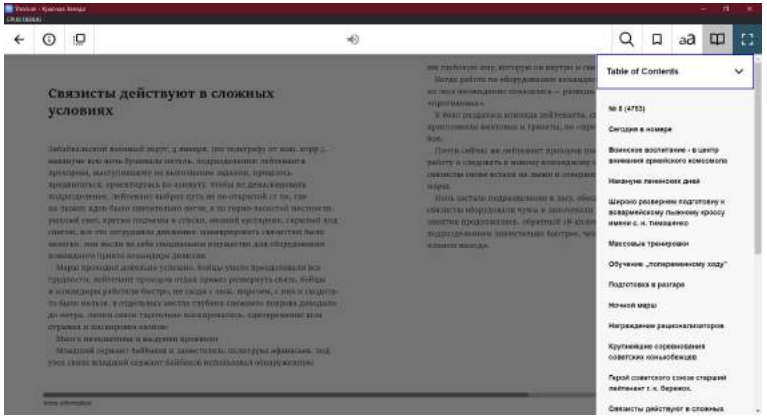
- HTML (совместимый с HOCR)
- EPUB3
- METS/ALTO (Экспериментальная поддержка)
- IIIF (Экспериментальная поддержка, разработка остановлена)
- Экспорт пакета с исходным содержимым контент-блоков и результатов анализа для внешних решений.
- Экспорт серии записей для каталогов.соответствующих стандарту OPDS2

«Прогресс» — апробация

В апреле 2020 года заработала пилотная техническая реализация, которая для подтверждения концепции, преобразовала более 1500 полос газеты «Красная звезда» за период 1941-1945 гг. в формат EPUB3, сформировнный в соответствии со всеми рекомендациями стандартизирующего органа. На момент апробации на этапе OCR использовался ансамбль сторонних OCR решений, на данный момент готовится к публикации внутренняя система для обучения СПО OCR решений, которая выводит их работу с архивной периодикой на необходимый нам уровень качества. Качество отработки всех остальных этапов на апробации значительно превысило аналоги.

Справедливости ради, аналогичная апробация на архиве изданий журнала “Советский спорт” оказалось провальной для нейросетевой подсистемы сегментации, “обученной” исключительно на образцах газет и после апробации система отправилась на доработку этой и ряда других подсистем





Каталог выпусков «Красной звезды» в формате OPDS2 и отдельный выпуск, открытые в EDR Lab Thorium Reader (Desktop)

Инженерия знаний, направленная на системы гибридного интеллекта

Если рассматривать ценность текущих цифровые и еще не оцифрованные фонды библиотек с прагматичной точки зрения, то по большей части они состоят из текстов, которые человек уже никогда не будет читать “от корки до корки”.

Фонды содержат колоссальные объемы информации, неравномерно полезной для улучшения работы автоматики и развития, качественного улучшения структуры человеческого знания.

Полезные для человека фрагменты информации, факты, выводы, идеи, художественно выдающиеся фрагменты разрежены и не доступны в массовом и персонифицированным, относительно индивида и его интересов, виде иначе, чем через средства автоматизированной экстракции, адаптации и навигации с AI-решениями в качестве ассистента, нахождение механизмов и драйверов для поддержания вовлеченности в процесс взаимодействия контента.

С другой стороны для реализации сценариев “ассистирования” от автоматизированных решений требуется высокое качество работы с естественным языком, чтобы сбалансировать существующее смещение в плане доступности информации в сторону технических специалистов.

Из этого вытекают направления:

- AI-Assisted UX
- Персонифицированная структуризация знаний
- Персонифицированная оптимизация формы представления знаний
- Специализированные направления оцифровки, ориентированное на “обучающий” материал для AI и NLP решениям
- Консолидация имеющегося языкового материала и стратегическое планирование работы с ним

Корпус темпоральных конструкций русского языка

В процессе поиска NER решений для изданий стало понятно, что критически важной функцией является детекция и интерпретация дат, интервалов, временных протяженностей и т.п., без понимания этих языковых структур автоматика не может делать выводы о причинно-следственных связях, формировать из полотна разрозненной фактологии конкретные истории, делать выводы об одной или разных одноименных сущностях идет речь.

Разработка решений для анализа подобных конструкций невозможно без инструментария проверки качества работы, в противном случае, не было бы понимание сколько конструкций не было выявлено или какие из них были интерпретированы некорректно. А результат любых изменений в логике анализа был бы слабо измерим.

Отталкиваясь от более богаты корпусов и методологий семантической разметки подобных конструкций мы сформировали базовый корпус для конструкций, которые описывают конкретные временные отметки с различной точностью, а также и их удаленность от заданного временного контекста (Например, 2020-18-02)

В целом это более 500 варьирующихся образцов с семикратной экспертной и не экспертной верификацией в машиночитаемой форме.

Корпус темпоральных конструкций русского языка - верификация и апробация парсеров

Огромную помощь в верификации начальной ревизии корпуса нам оказали учащиеся образовательного проекта Школа 21, под эгидой ПАО Сбербанк, в рамках тематического хакатона.

Они многократно перепроверили датасет и апробировали на нем полторы дюжины различных подходов к логике парсинга. Абсолютно искренне мы хотим отметить, что учащиеся проекта “Школа 21” сильно выделяются, как своим уровнем профессиональной мотивации, так и готовностью работать с новыми для них доменами, на общем фоне IT специалистов, присутствующих на локальном рынке труда или вовлеченных в решению типовых задач на местных предприятиях. Еще серьезнее он отличается от результатов стандартной ВУЗ-овской подготовки. Хочется подчеркнуть, что речь идет не о использовании некого нового источника “бесплатной рабочей силы”, а о том, что мы видим смысл предлагать выпускникам лучшее из того, что мы в состоянии предложить в плане компенсационного пакета и задач.

Задача оказалась предельно сложной, автоматика приблизилась только к 40% корректно интерпретированных конструкций, группы, занимающиеся верификацией смогли корректно интерпретировать чуть более 70% конструкций, Более 10% конструкций были идентично некорректно интерпретированы при тройной кросс-проверке.

Текущая ревизия корпуса, дополнительные материалы и различные связанные статистики доступны по адресу https://github.com/ndlrf-rnd/school21-retropress-temporal

Корпус темпоральных конструкций русского языка - проблема семантической разметки корпуса

Одним из конечных итогов мы видели интеграцию ISO TimeML/Timex3,в корпус стандартов ГОСТ, а также адаптировать гайдлайны по применению этой разметки для русского языка. Но столкнулись с проблемами, хорошо описанными специалистами РАН, работающими над аналогичной задачей (http://www.mathnet.ru/links/1d949476b4e12dbf4818288adfb05045/ps241.pdf)/

В данный момент мы сделали ставку на переводове направление французской лингвистики: анализ микросинтаксиса (о том, что такое микросинтаксис можно почитать тут: http://iitp.ru/upload/publications/7304/05%20Iomdin.pdf) Микросинтаксис похож на очень значимый, но представлявший большие трудности при анализе традиционным способом, уровень, который хорошо отражает “семантико-лексическую” механику, и выглядит оптимальным для формализации логики нанесения семантической разметки экспертами, а также задания конкретных правил детекции и интерпретации для автоматики.

Так как уровень и степень локализации микросинтаксических конструкций крайне схож со структурами выявленными путем абляционного анализа языковых моделей класса “Transformer” (https://pair-code.github.io/interpretability/bert-tree/), рабочей гипотезой является отделение и анализ всех “отзывчивых” на текморальные структуры нейронных ансамблей обученных трансформеров, Тенденция этих структур к гармонизации отображения в геометрии с отрицательной кривизной, некотору.ю кэспертизу по которым мы смогли получить в процессе работы над консолидацией графов знаний, позволяет надеяться на принципп “fail fast”, когда нежизнеспособность гипотезы дает о себе знать ранее, чем в работу вкладываются значительные ресурсы.

Оцифровка и формирование коллекций для AI/ML

В качестве инициативы ОИЗ НЭБ, первыми в мире мы начали пилотное направление по **оцифровке изданий**, для которого отбор изданий ориентирован **не на человека-читателя, а на обучение AI/ML решений**.

Отбираются изначально “дружелюбные” и схожие с концепцией “Датасета” формы материала:

- Иллюстрированные словари
- Словари
- Справочники
- Картографические и конструктивные альманахи
- Индексы
- Книги с ребусами и детскими раскрасками
- Тексты стандартов, инструкций, регламентов
- Любые издания с табулярными данными.

Пилотный блок оцифрованных (без сегментации и аналитического слоя) изданий, отбор которых осуществлен экспертно специалистами ОФК НЭБ, доступен в НЭБ: https://docs.google.com/spreadsheets/d/1pzqLtUCx-G4a8v07RLfaGO0TlE7gCEdt9zq3lKY6oLA/edit?usp=sharing основная задача, на которую направлен пилотный блок - Visual Grounding и Visual grounded language communication, которая позволяет убрать деструктивный фактор языкового дрифта, которому подвержены модели автоматических коммуникационных агентов.

В планах автоматизированный классификатор для отбора подобных изданий из каталога НЭБ.

Автосистематизация документов - проблематика

В рамках работы над новым разделом НЭБ - Наука со стороны ОИЗ РГБ была разработана концепция автоматической систематизации (рубрикации) материала. А также модель навигации пользователя-человека по большим банкам текстовых документов. Смежно-вытекающими функциями является возможность использования решения в качестве рекомендательно-экспертной системы.

Был разработан набор прототипных программных решений, близких по характеристикам к эксплуатационному сценарию.

Ключевым нововведением является решение основной проблемой классификационных систем, делающей якобы “невозможным” формирование и использование их как единой системы является зависимость иерархии терминов и концептов от контекста конкретной задачи или специализации пользователя. “Статистические методы в биологии” могут быть как частной прикладной областью для статистических методов, так и одним из наборов аналитических аппаратов в биологии. Это противоречит удобной для навигации и формирования древовидной структуре классификационной системы.

Несмотря на привлекательности идеи просто формировать наборы ключевых слов на основе семантики документов, этот подход не позволяет однозначным и полезным образом утилизовать существующий огромный объем результатов экспертной работы по рубрикации и прочей высокоуровневой систематизации знаний.

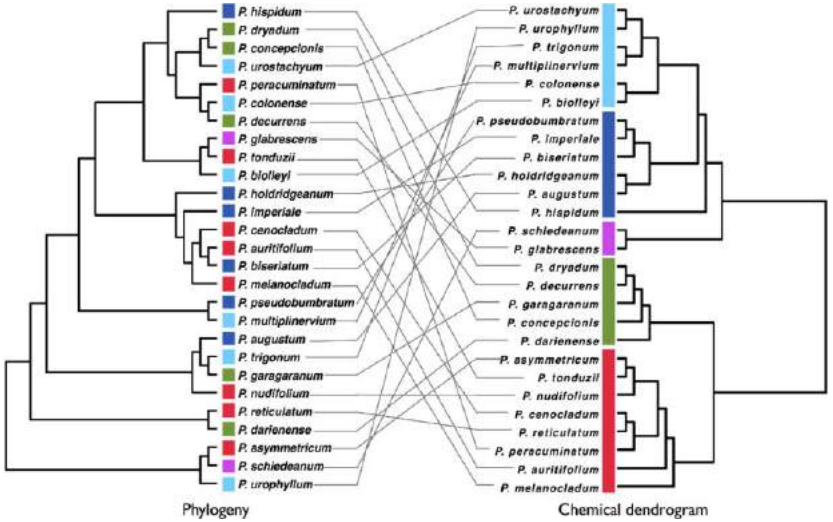


Иллюстрация взята из статьи: Salazar, Diego & Jaramillo, Alejandra & Marquis, Robert. (2016). Chemical similarity and local community assembly in the species rich tropical genus Piper. Ecology. 97. 3176–3183. 10.1002/ecy.1536.

Автосистематизация документов - решение

В качестве основного подхода была выбрана консолидация графов знаний в геометрии пространств низкой размерности с отрицательной кривизной. Это позволило совместить различные классификационные системы, описывающие области знаний, тематики и т.п. Для консолидированного представления осуществлялось геометрическое центрирование и обратная проекция в удобные для навигации древовидные структуры, в зависимости от заданного контекста (например, специализации исследователя, работающего с системой).

Привязка документов и их семантически однородных текстовых фрагментов осуществляется через векторизацию текста авторегресионными языковыми моделями. Выбор имплементации вычислительных методов осуществлен между и Facebook Poincaré embeddings for learning hierarchical representations и Stanford Hazy Research Hyperbolic в пользу последней, так как решение от Facebook не имело механизма детерминированной проверки результатов работы стохастических методов.

- Краткое иллюстрированное описание подхода доступно по ссылке: https://docs.google.com/presentation/d/1c3Wc2YOteupvLNdRQeoALxdTN3He52z1FwmuwNG-_Ek/edit?usp=sharing
- Рабочее концептуальное описание проблематики и технологии: https://drive.google.com/file/d/11KgqaGeqv8Jq7foFmGG2La2YtDN5TjSj/view?usp=sharing
- Техническое описание архитектуры решения и подходов к выбору языковых моделей и подсистем публично размещены по адресу: https://github.com/ndlrf-rnd/systemize-doc (формат: MarkDown)
- Исходный код системы планируется к публичному размещению после завершения внутренней апробации и экспертного ассессмента логики работы решения

Автосистематизация документов - текущий статус

В данный момент у нас работает несколько вариантов опытных образцов авто-классификаторов, помогающих моментально привязать доступный фрагмент текста к системе ББК, УДК, классификат ору специальностей ВАК и т. п. Апробируются как различные методы геометрического встраивания, так и механизмы, позволяющие системе следовать не только общему пониманию смысла языковых конструкций, но также и логике экспертной группы, при использовании даже небольшого количества данных о работе этой группы.

Отдельно исследуется возможность гибридного применения пространств продуктов (product manifolds), по гипотезе, позволяющие разумным для человека образом разрешать и предотвращать сложные терминологически-понятийные конфликты в процессе классификации или навигации по банкам данных.

Затрудняющим дальнейшее развитие технологии фактором является необходимость проведения ресурсоемких вычислительных экспериментов для сопоставления вариантов реализации с разными языковыми моделями и разными подходами к построению геометрии семантических пространств. Качество результатов вычислительных экспериментов значительно зависят от точности операций с FPU числами.

Огромное преимущество могут дать решения на базе “Эльбрус”, поддерживающие на аппаратном уровне векторизованные операции над 128-битными FPU (В два раза больше, чем в случае решений с DPTC от компании NVIDIA)

Раскрытие человеческого и кадрового потенциала

Большое количество актуальных проблем находится в гуманитарной области, в которой технические специалисты склонны проявлять массовый профанизм.

Носителями гуманитарной экспертизы являются специалисты, для которых крайне высоким может являться интерфейсный барьер специализированных систем.

Дружелюбные интерфейсы и единое пространство для работы гуманитарных экспертов является критическим звеном в решении задач по построению Единого российского цифрового пространства знаний (ЕРПЗ)

Проблемы и решения в понимании и реализации ОИЗ НЭБ, связаны этой темой.

Полнокорпусная лингвистика - Проблематика

12-13 декабря 2020 года сотрудники R&D НЭБ приняли активное участие в DH Advent 2020: зум-митап Центра цифровых гуманитарных исследований ВШЭ, что послужило началом академического взаимодействия.

Проблематика, обозначенная участниками была проанализирована на предмет возможности некой явной стратегии решения таких проблем как:

1. Сложность измерения вклада гуманитарных специалистов и обоснования ценности их экспертизы на рынке труда
2. Репрезентативность и качество русскоязычных текстовых корпусов
3. Юридические блокеры в доступности и развитии Корпуса русского языка
4. Унификация пространства работы над корпусами и прочими языковыми датасетами
5. Проблема доминирования authority institution downstream модели в формировании и разметке корпусов, отсутствие управляемого процесса включения community upstream contribution.

Для нас эти проблемы выглядят блокирующим фактором в развитии автоматики, взаимодействующей с людьми посредством естественного русского языка.

Полнокорпусная лингвистика - Концепция решения

В качестве решения мы предлагаем рабочий вариант стратегии “Полнокорпусной лингвистики”, которая ставит своей целью расширение границ корпуса русского языка до всех технически доступных текстовых и прочих данных, что по большей части снимает проблему репрезентативности в широком смысле. Второй целью является включение в процесс структуризации и разметки “Полного корпуса” всех потенциально значимых специалистов в языковых, исторических и прочих гуманитарных областях, с некими наглядными и общими правилами оценки их вклада.

Фактор существующих нормативно-правовых блокеров возможно почти полностью сократить в случае формирования рабочего пространства, которое организационно и технически функционирует в рамках протекции ФЗ “О библиотечном деле”, “ФЗ О национальной электронной библиотеке” и ФЗ “Об обязательном экземпляре”. В совокупности они дают возможность в рамках библиотечной экосистемы вести работу специалистов фактически с любыми технически доступными текстовыми данными в исследовательских целях.

Планирование работ по экспертной разметке и структуризации языковых и высокоуровневых смысловых конструкций предполагается как жизненный цикл экспертно размеченных данных в рамках подхода Weak Supervision. Конкретные механизмы на данный момент активно прорабатываются, **к участию приглашаются все потенциально заинтересованные организации**.

Предложенная концепция видится крайне предметным способом решения всех обозначенных на DH Advent Meetup выше проблем, с заложенной в основу положительной обратной связью, мультиплицирующей вложенные усилия. А также базой для осуществления **беспрецедентного скачка в качестве работы NLP AI систем,с русским языком до уровня** аналогичного современным крайне существенным достижениям в области **англоязычной и китайской корпусно-дистрибутивной лингвистики**.

Подробное описание концептуального видения доступно по ссылке:
https://docs.google.com/presentation/d/1DWd-Hr7SpMmdoF3GqQ0fa1ezkpYGFr1OHHyHIM4lir4/edit?usp=sharing

Полнокорпусная лингвистика - PoC

В качестве практического обоснования R&D НЭБ собрали из доступных источников, “технически очистили” и языковую атрибуцию для корпуса из **более чем 20 000 000 000 русских слов**. На данный момент ведется работа по временной атрибуции строк с точностью до декады.

Потенциально-целевой объем “полного” корпуса предполагается большим, чем на один порядок.

Трансфер результатов направления “Полнокорпусной лингвистики” в публично доступное пространство предполагается в виде:

- Агрегатов детальной описательной статистики
- Частотных словарей
- N-gramm моделей
- Статистических поправок для существующих публично доступных корпусов
- Обученных в “чистой зоне” NLP/ML моделей при условии отсутствия эффекта “дословного запоминания” данных.
- Наборов строк преобразованных так, чтобы при сохранении своих совокупных статистических свойств

Процесс обучения автоматики, со включением данных традиционных экспертных заключений

На примере подсистемы оцифровки, осуществляющей сегментацию документов мы обнаружили массовые дефекты в широко используемых для этого наборах данных от PriMa Lab и других поставщиков.

В качестве решения мы попросили эксперта рассмотреть и написать компактное заключение для ~500 образцов обучающей выборки, имеющей формальную разметку в форматах Page XML и METS в любой удобной для него форме.

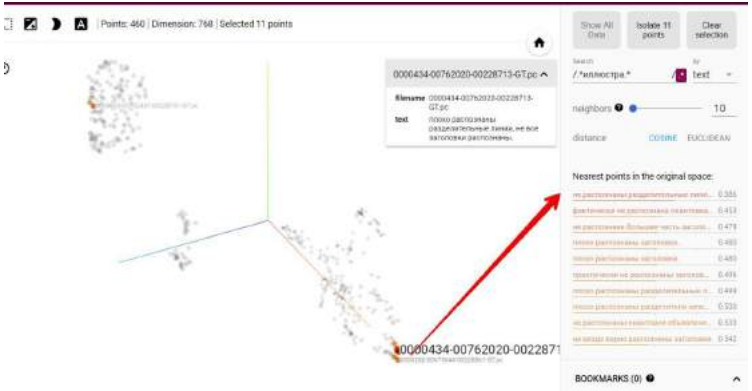
Тексты заключений были векторизованны через языковую модель, обучаемая система (класс U-Net+) была изменена так, чтобы получать на вход и размеченные образцы и вектор экспертного заключения об их качестве и характере найденных дефектов.

В процессе обучения автоматика смогла успешно использовать экспертные заключения для коррекции степени влияния дефектных образцов на результирующее качество. Содержимое заключений также по природе оказалось хорошо кластеризуемыми с целью нахождения типовых проблем в образцах golden set

В результате для значительного улучшения качества работы ML решения эксперту-человеку не пришлось обучаться работе со специализированным ПО для разметки, а также в чем либо отходить от привычной для него формы работы.

Мы не видимо ограничений для массового использования и масштабирования найденного подхода.

Исходные данные с экспертными заключениями по датасету PRImA Layout analysis dataset доступны по ссылке: https://gist.github.com/mrjj/439b7c096af1eace1a04ee19803bf7ac



Замкнутый цикл, когда на основе заключений выделяются типовые проблемы, после автоматизированной сверки с которыми автоматика дозапрашивает экспертные заключения для потенциально проблемных образцов видится процессом с положительной самоамплификацией относительно обеспечения качества обучающего материала, что органично дополняет решения проблемы скрытой стратификации данных выделенной в качестве наиболее актуальной и значимой специалистами Stanford University Hazy Research Group (https://arxiv.org/abs/1909.12475)

Решения по “достаточной” автосуммаризации

Специалистами R&D НЭБ был осуществлен ряд PoC экспериментов с целью сокращения объема механического чтения документа минимально деструктивным для смысла способом.

Данный подход противоположен традиционному подходу к сокращению объема контента до заданного фиксированного объема.

И его задача - убрать менее значимую информацию из целевого объема текста ровно настолько, сколько нужно, чтобы степени вовлеченности человека хватило для потребить всего потенциально значимого объема материала. Можно назвать этот подход “прогрессивной суммаризацией” или представить его как задачу ранжирования по релевантности интересам человека как на уровне документов, так и фрагментов внутри отдельных документов.

В данный момент технология получила интерес со стороны российского коммерческого IT сектора, ставящего своей целью глобальный англоязычный и китайский рынки профильных SaaS/PaaS решений.

Более детальное описание проблематики и рассмотрение подходов к решению проблемы доступны в виде рабочего документа “*What does TL;DR mean? (TL;DR - Too long didn’t read)*” по ссылке:
https://docs.google.com/presentation/d/1LrvHcb771WI4AgUP2JWk0lDvCDfV2RZ0z9c7iu3GFn4/edit?usp=sharing

Проблема опосредованности экспертного знания

Экспертное знание, выражаемое через экспертное заключение можно представить как в значительной мере обусловленное знакомством эксперта с неким конечным количеством образцов, дающих представление о предмете заключения и характере его вариативности. Также, использование конкретных примеров предмета заключения и негативных образцов того, что предметом не является (“Дельфин не является рыбой, несмотря на значительное таксономическое соответствие группе рыб”)

Рабочей гипотезой о причинах ситуаций с отсутствием консолидации экспертного кворума, как и проблема его субоптимальной консолидации является двойная природа работы эксперта - ознакомленность с образцами предмета и аналитическая составляющая, дающая возможность эксперту генерализовать и экстраполировать знание о конкретных образцах. Если изолировать эти факторы, можно существенно снизить фактор “черного ящика” в работе экспертов.

Наиболее реалистичным механизмом подобной изоляции является предметизация образцов, известных эксперту. Но так как доступность всех знакомых конкретному эксперту материалов вряд ли возможна, как и корректность и полнота человеческой памяти относительно них, решение выглядит как отбор экспертом образцов из большого доступного набора с быстрой наглядной обратной связью, дают ли эти образцы ожидаемую экспертом картину дифференциации всего набора. Отличием от Supervised ML методик обучения классификаторов заключается в отзывчивости петли обратной связи.

Предметизация экспертного знания экспериментальный дизайн

В рамках проекта “Книжные памятники” была осуществлена векторизация около 1 000 000 сканов страниц средствами кодирующей модели VGG19, предобученной на образцах ImageNet. Итоговые векторные представления были размещены в единое индексное пространство под управлением системы NGT от компании Yahoo! Japan (компании предоставлена обратная связь, есть планы по совместной доработке решения с целью эффективного построения и эксплуатации индекса на объемах записей существенно превышающий текущий целевой размер в 10 000 000 000 векторов)

Экспертам было предложено интерфейс, дающий возможность отбирать положительные и негативные образцы с целью определения определенных функциональных и стилистических классов графических элементов.

Система отдает срез ранжированных результатов в soft-realtime режиме в ответ на добавление или исключение образцов.

Первая версия системы доступна по ссылке: https://storage.rusneb.ru/_kspace/# (Известным дефектом первой версии является неадекватная чувствительность к негативным образцам)

Предметизация экспертного знания
итоги эксперимента

Высокая скорость и крайне незначительное количество необходимых образцов показало использованный UI механизм как очень эффективный и точный способ задания целевого среза для векторизованных данных в случае, если он известен (“схематические изображения стрелковых орудий в технике линогравюры” или “библиотечные штампы XIX века”), а также механизм нахождения не ожидаемых экспертом закономерностей и значимых срезов/таксономий.

Экспертный отбор “характерных образчиков” через дифференциальный механизм использованный на фиксированном датасете (при наличии определенных методик балансировки) видится, как потенциально очень компактный и простой для формирования, но при этом высокоселективный и гранулярный способ выражения экспертного знания, удобный для использования при обучении классификаторов в рамках Weak supervision подхода, а также анализа природы и разрешения противоречий в экспертных заключениях. Фактически, экспертное мнение оказывается тождественным некой пиктоидеограмме, составленной из определенного “алфавита” образцов.

Очевидной становится механика атрибуции экспертов как носителей консолидированного ядра понимания предмета, так и представителей маргинализированных форм понимания, что дает принципиально новый уровень контроля при формировании экспертных групп и советов.

По ресурсовзатратом этот способ несопоставим с традиционной “разметкой датасетов”. Также, решение в теории может быть наименее ресурсоемким способом предметизации высокоуровневых языковых абстракций таких как классы, категории, концепты.

Решение полностью применимо и в качестве поискового механизма для казуального (не-экспертного пользователя). Решение также имеет потенциал, как механизм обеспечения engagement при знакомстве с историко-архивных материалов, источник significant observations flow для формирования высокоуровневых академических концепций.