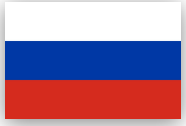
Комплекс программных решений и рабочий процесс по оцифровке изданий

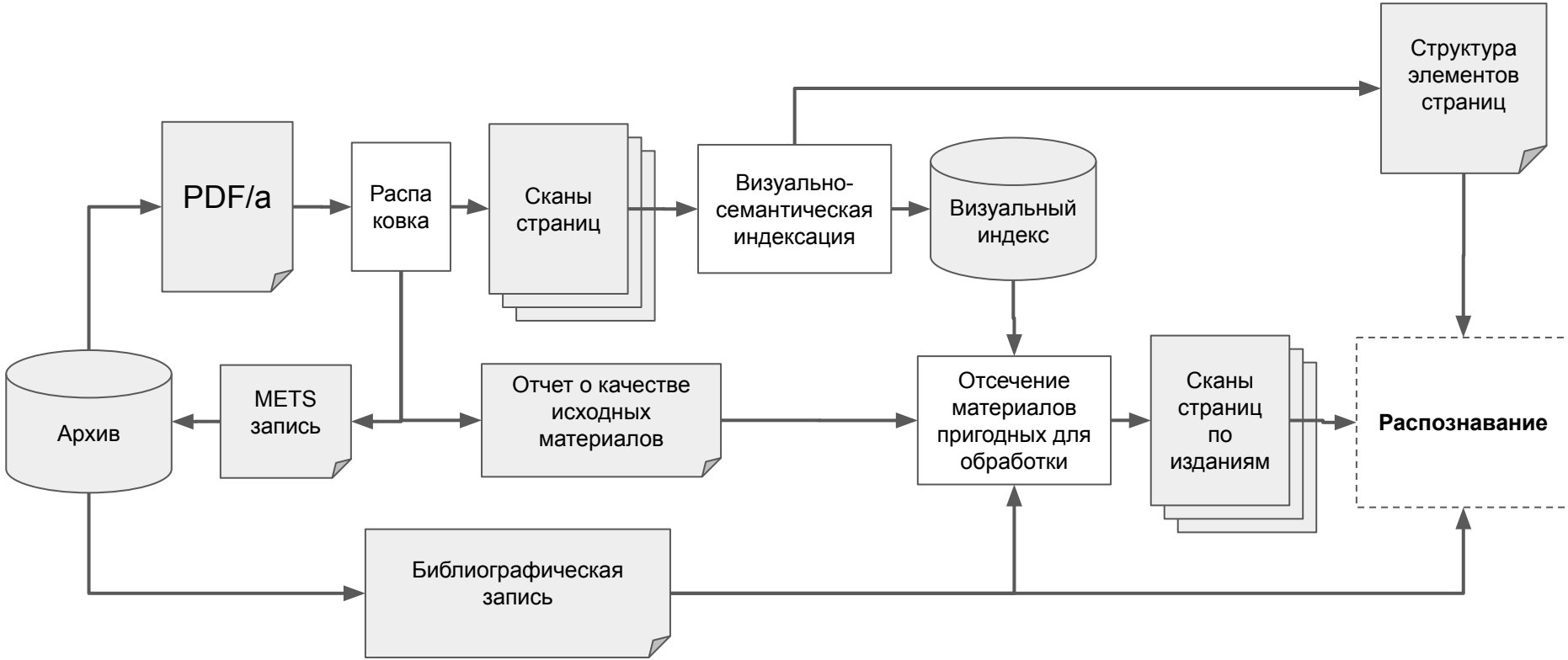
РГБ ЛИР/ОИЗ, февраль 2020

Проектные особенности:

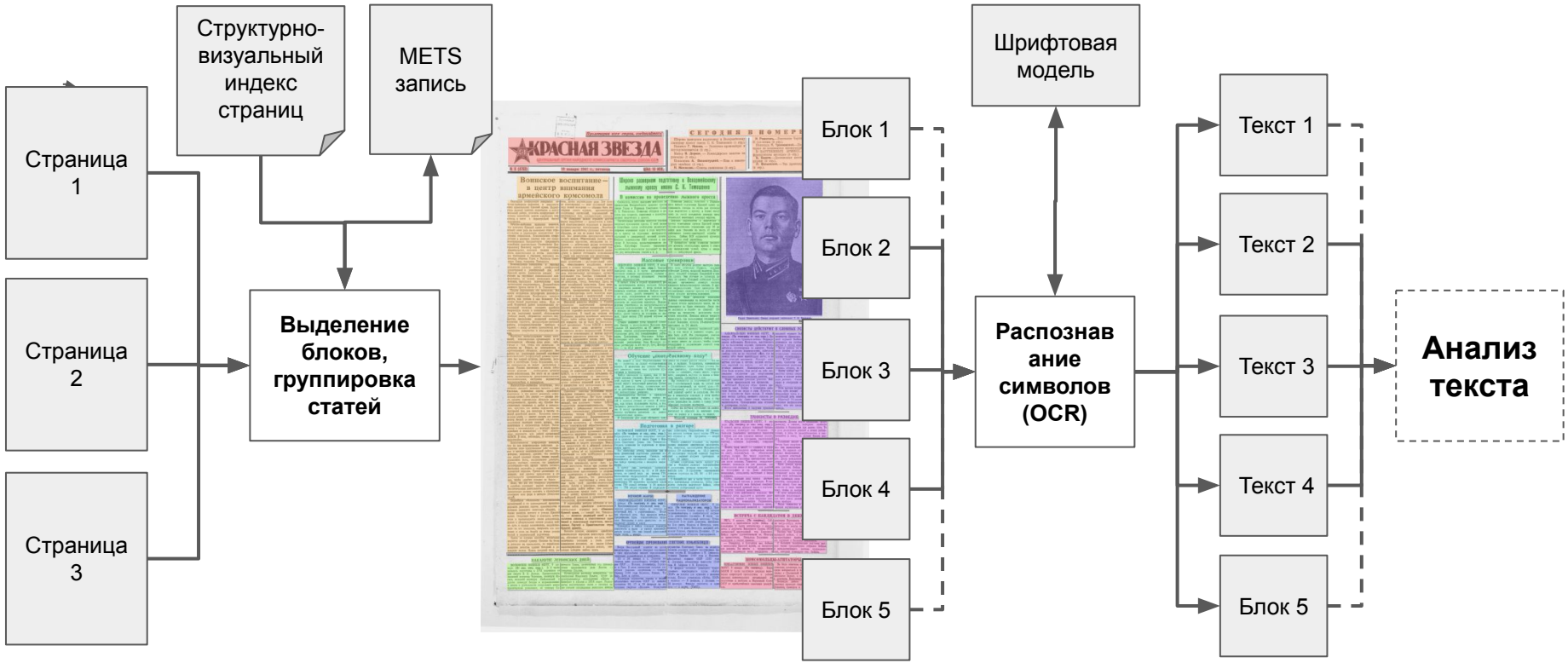


- Русский язык:
  - **Флективная морфология**.
  - Множество лексических изменений на протяжении XX века.
  - Уникальные термины и конструкции для временных данных, имена собственные.
  - Топонимы из других языков.
  - Необязательная диакритика.
  - Правовые аспекты связанные с основными корпусами, ухудшение уровня поддержки русского языка у зарубежных систем.
- Русские издания XX века:
  - Большое разнообразие типографских процессов.
  - Большое разнообразие и разброс качества гарнитур шрифтов и производства носителя.
  - Множество аутентичных художественно-декоративных стилей и традиций.
  - Большое количество “политинформ” лексики в периодике XX века.

Оцифровка документов - стадия отбора

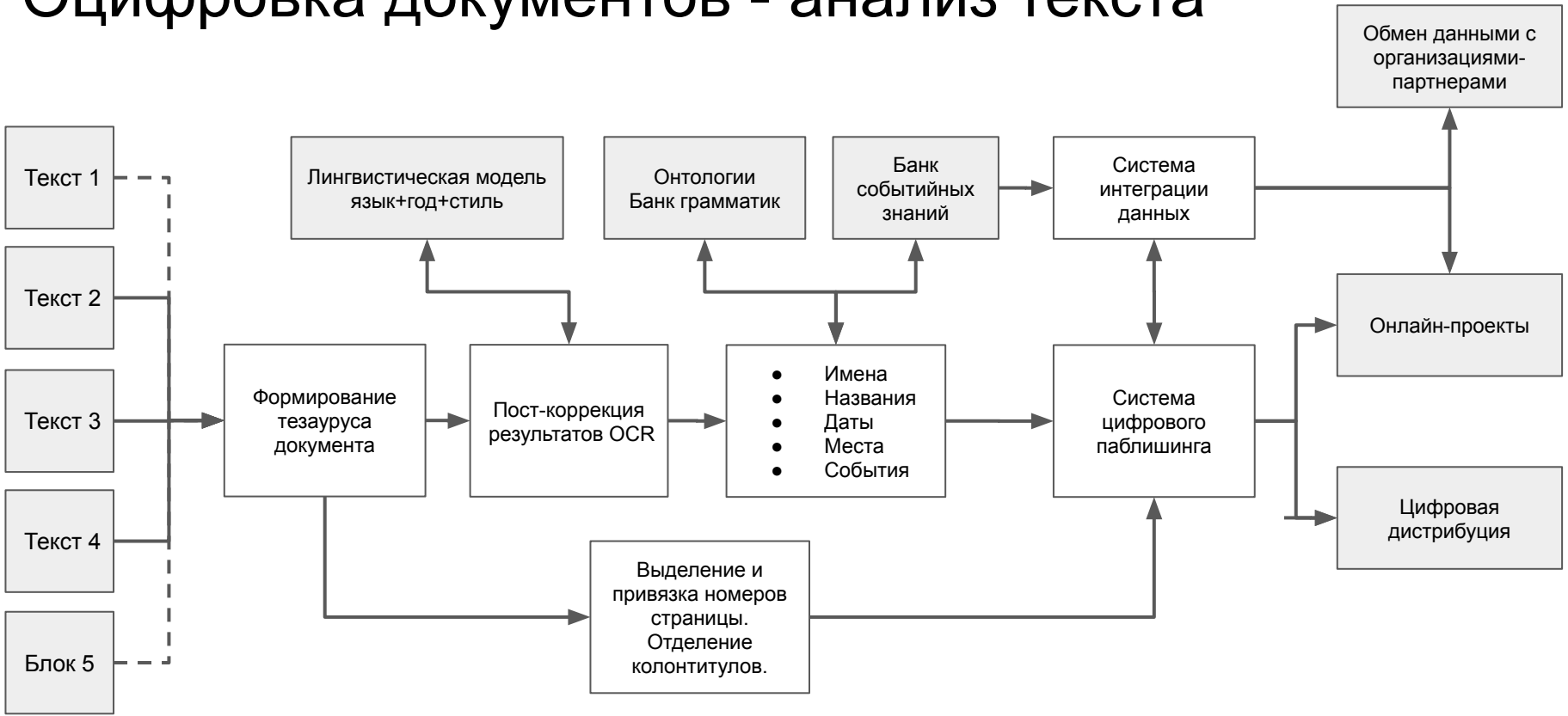


Оцифровка документов - распознавание



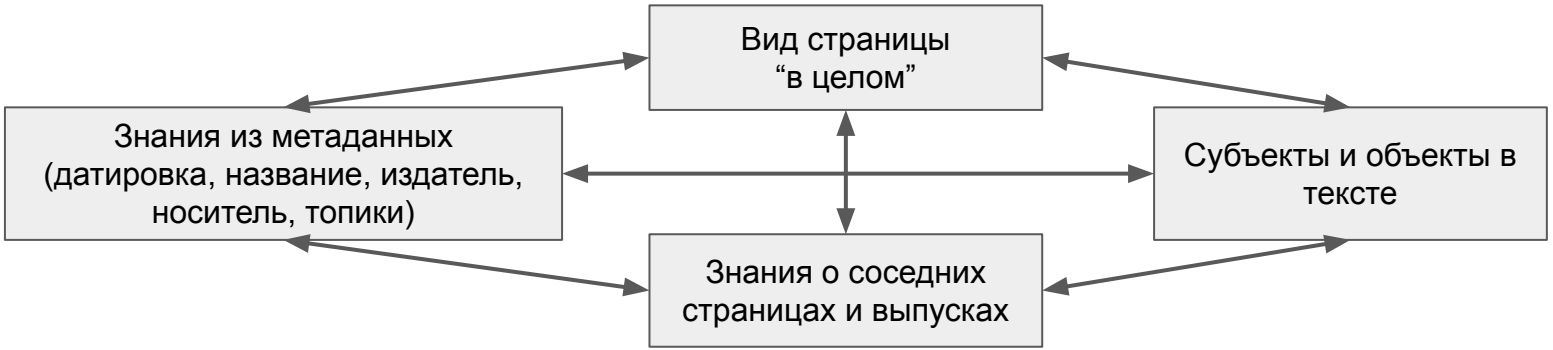
* В качестве примера была выбрана фронтовая газета “**Красная звезда**”, являющаяся предельно сложным случаем из-за особенностей типографики и лексики.

Оцифровка документов - анализ текста



Примечания по архитектуре:

- Модульная архитектура и популярные интерфейсы интеграции (SpaCY, Tensorflow, PyTorch/TorchVision, Page.xml, METS) предполагает заменяемость конкретных подсистем и нетребовательность к квалификации при обслуживании работы.
- Дизайн системы позволяет за счет одного уровня знаний об издании улучшать другой уровень:



Внутренние интеграции - визуальный индекс

Система на базе нейросетей и больших графов, которая видит, запоминает и сравнивает графический материал во многом похоже на то, как это делает человек используя ассоциации с известным и виденным ранее..

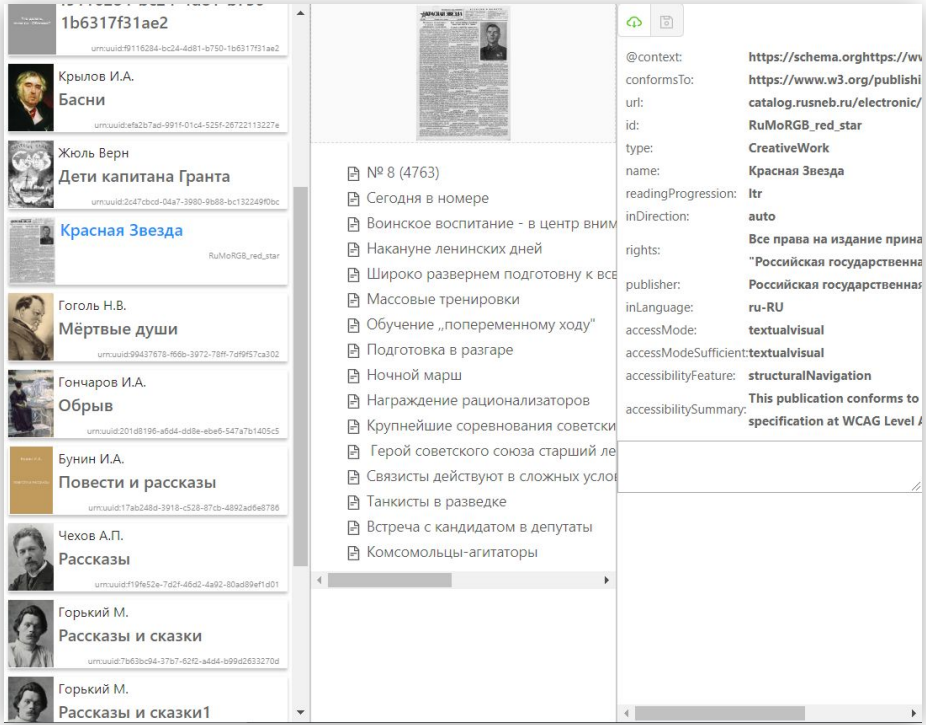
Наша реализация умеет работать с беспрецедентно большими индексами (более 10 млрд страниц) не теряя при этом в качестве.

Аналоги:

- Europeana Siamese
- BSB image search
- Pinterest
- Yandex - поиск по картинкам.

На иллюстрации системе “показывают” несколько примеров, того, что является первой и не первой страницей газеты “Красная звезда”, после чего система самостоятельно разделяет подшивку на выпуски.

Внутренние интеграции - NEB Pub Maker



Система позволяет скомпоновать результат оцифровки в отдельное издание, внести редакционные правки, дополнить метаданные.

Система автоматически готовит результат к публикации в форматах:

- EPUB3 + Structural semantic markup
- W3C WebPub
- Метаданные преобразуются к большинству распространенных форматов автоматически.

Место оператора - сторонние инструменты

PRImA Lab **Aletheia** - Проверка результатов анализа, создание новых и коррекция имеющихся образцов разметки.

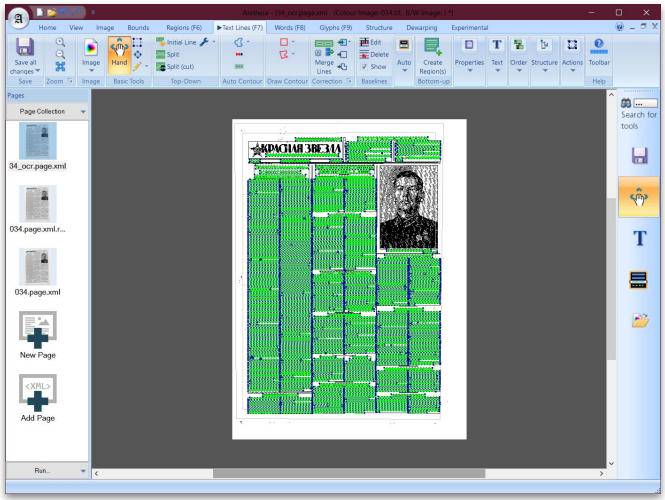


photo credit https://www.primaresearch.org/tools/Aletheia

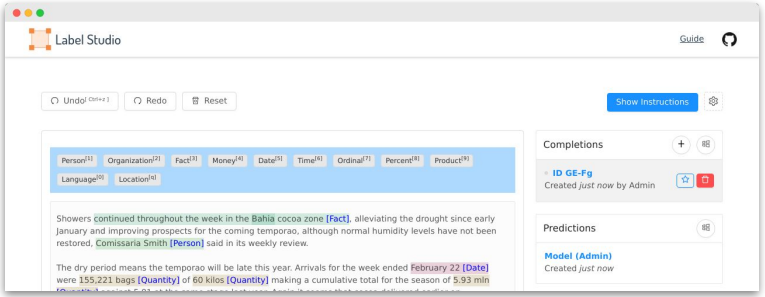


photo credit: https://www.heartex.ai/

Label Studio и **SpaCy** -
Разметка, коррекция гипотез в тексте, работы моделей определения сущностей

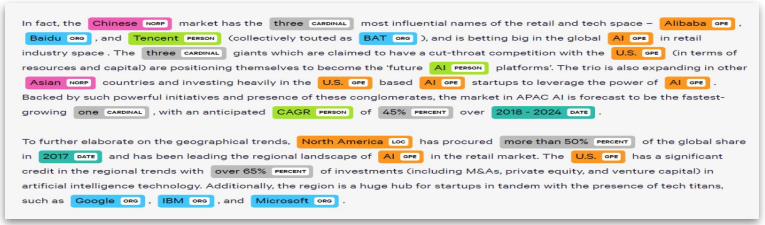


photo credit: meenavyas

ЛИР / Оцифровка документов / технологии и партнеры

made with europeana

europeana newspapers

Europeana newspapers
Luxembourg library Open data
Размеченные датасеты изданий.

Bibliothèque nationale du Luxembourg
Open Data

University of Salford MANCHESTER

PRImA Lab
Методики и инструментарий контроля качества.

BSB Bayerische StaatsBibliothek
Information for Excellence

Подходы к систематизации и организации работ,
технологии определения визуальной идентичности.

TECHNISCHE UNIVERSITÄT DARMSTADT

Ubiquitous knowledge processing Lab
Компоненты систем и алгоритмы структурного анализа знаний

UNIVERSITÄT LEIPZIG

Дизайн системы “потока страниц”, в которой визуальные и текстовые аспекты взаимно улучшают результат анализа.

EPFL

CEDILOR

Digital Humanities Lab
(Digital Humanities project) пост-обработка структуры документа.

YAHOO! JAPAN

Технологии индексации смысла больших объемов документов.
Индекс для экспериментального графового анализа порядка чтения.

Tesseract OCR

spaCy

Платформы для интеграции вычислительных лингвистических и AI моделей.

Яндекс

MyStem - стеммер русского языка.
Tomita parser - механизм фактологических грамматик

Поддерживаемые форматы и стандарты:

- Schema.org
- METS
- PRImA Page.XML
- HOCR
- FRXML
- EPUB3 structural semantic vocabulary