

Цифровая трансформация документов национальной системы стандартизации на основе подхода управления требованиями

ИИ и цифровая трансформация стандартов. Как переводить нормы
в цифровые требования без потери смысла и контекста.

Сергей Трофимов

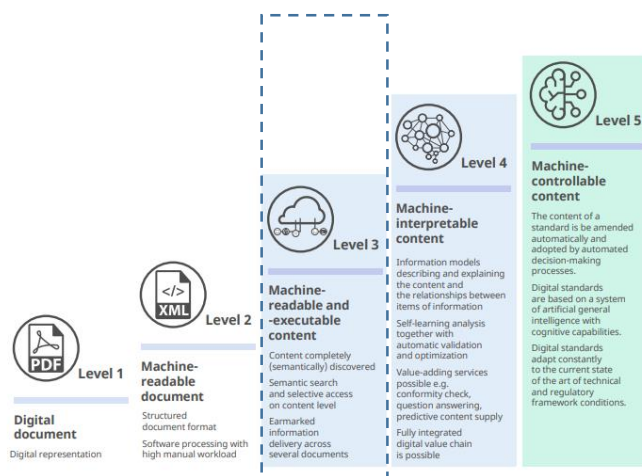
| советник генерального директора
Российского института стандартизации



ЦИФРОВЫЕ СТАНДАРТЫ. КЛАССИФИКАЦИЯ

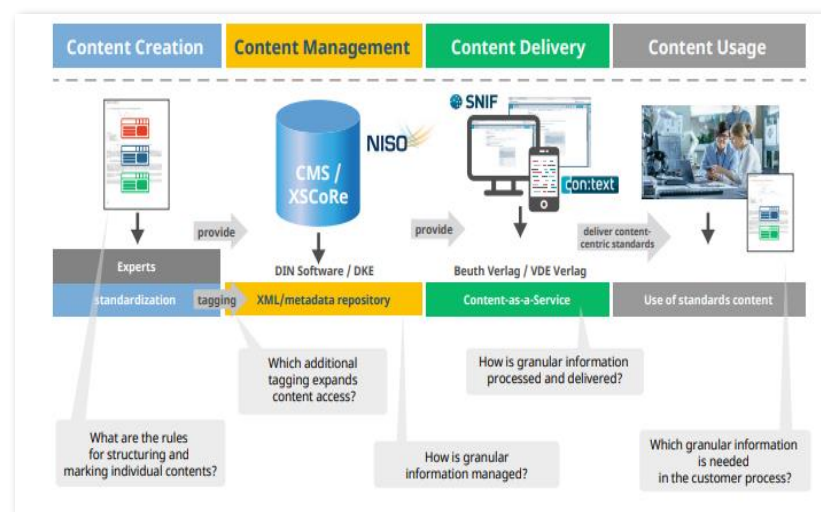
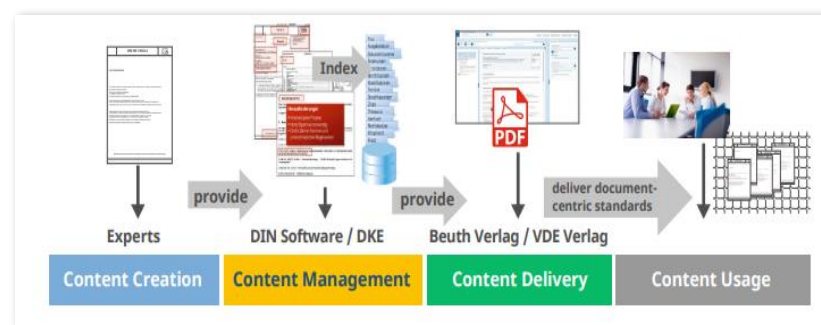


ПОДХОДЫ DIN И DKE

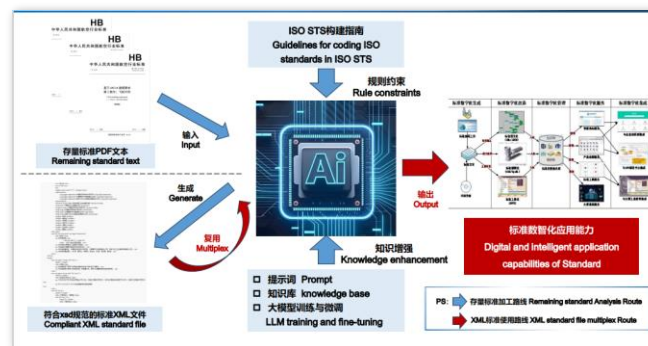
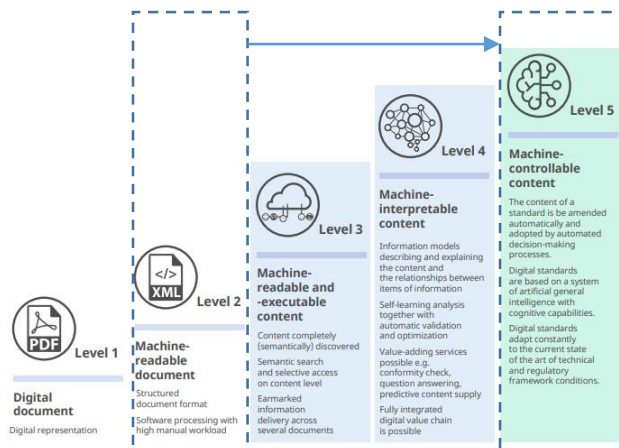


1) NISO STS – набор тэгов и атрибутов (593 шт.), основа Fonto (редактор OSD платформы ISO и IEC). Все стандарты ISO и IEC разрабатываются имея XML основание

2) SNIF - это **система для семантической индексации стандартов**. Она связывает интеллектуально разработанные метаданные с семантически проиндексированными текстами стандартов, чтобы предоставлять более качественную информацию системам, участвующим в дальнейшей обработке.



ПОДХОДЫ AVIC CHINA AERO-POLYTECHNOLOGY ESTABLISHMENT



- 1) Используют NISO STS
- 2) Подключают LLM модель и переобучают, для получения XML семантики и дальнейшего применения в различных информационных системах (PLM, ERP, CAD) – используя перевод в различные форматы (json, reqIF и пр.)

标准原文 Original standard

表格元素 table element

尺寸	公差	公差带	公差带	公差带	公差带	公差带	公差带	公差带	公差带
1	10	±0.1	10	±0.1	10	±0.1	10	±0.1	10
2	10	±0.1	10	±0.1	10	±0.1	10	±0.1	10
3	10	±0.1	10	±0.1	10	±0.1	10	±0.1	10
4	10	±0.1	10	±0.1	10	±0.1	10	±0.1	10
5	10	±0.1	10	±0.1	10	±0.1	10	±0.1	10

XML示例 Standard XML Example

```

<table>
  <tbl_struct>
    <tbl_header>
      <tbl_header_row>
        <tbl_header_cell>
          尺寸
        </tbl_header_cell>
        <tbl_header_cell>
          公差
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
      </tbl_header_row>
    </tbl_header>
    <tbl_info cols="10">
    </tbl_info>
    <tbl_r cells="10" ix="1" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="2" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="3" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="4" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="5" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
  </tbl_struct>
  <tbl_r cells="10" ix="1" maxcspan="1" maxrspan="1" usedcols="10">
    1
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="2" maxcspan="1" maxrspan="1" usedcols="10">
    2
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="3" maxcspan="1" maxrspan="1" usedcols="10">
    3
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="4" maxcspan="1" maxrspan="1" usedcols="10">
    4
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="5" maxcspan="1" maxrspan="1" usedcols="10">
    5
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
</table>

```

LLM解析结果 Analysis results of LLM

```

<table>
  <tbl_struct>
    <tbl_header>
      <tbl_header_row>
        <tbl_header_cell>
          尺寸
        </tbl_header_cell>
        <tbl_header_cell>
          公差
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
        <tbl_header_cell>
          公差带
        </tbl_header_cell>
      </tbl_header_row>
    </tbl_header>
    <tbl_info cols="10">
    </tbl_info>
    <tbl_r cells="10" ix="1" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="2" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="3" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="4" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
    <tbl_r cells="10" ix="5" maxcspan="1" maxrspan="1" usedcols="10">
    </tbl_r>
  </tbl_struct>
  <tbl_r cells="10" ix="1" maxcspan="1" maxrspan="1" usedcols="10">
    1
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="2" maxcspan="1" maxrspan="1" usedcols="10">
    2
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="3" maxcspan="1" maxrspan="1" usedcols="10">
    3
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="4" maxcspan="1" maxrspan="1" usedcols="10">
    4
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
  <tbl_r cells="10" ix="5" maxcspan="1" maxrspan="1" usedcols="10">
    5
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
    10
    ±0.1
  </tbl_r>
</table>

```

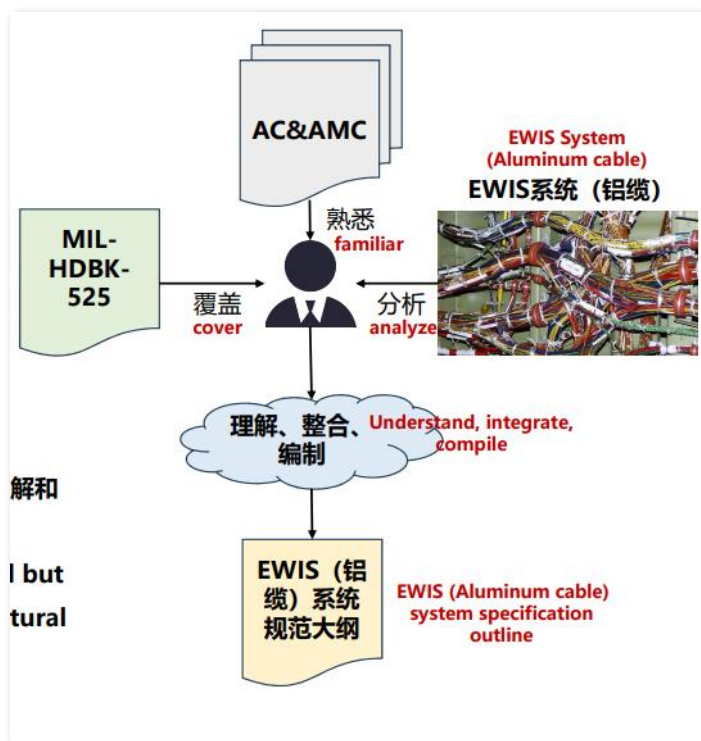
表格能被识别、定位、编码
Table can be recognized, located, and encoded

表格按照行列顺序记录
Table is recorded with sequence of rows and columns.

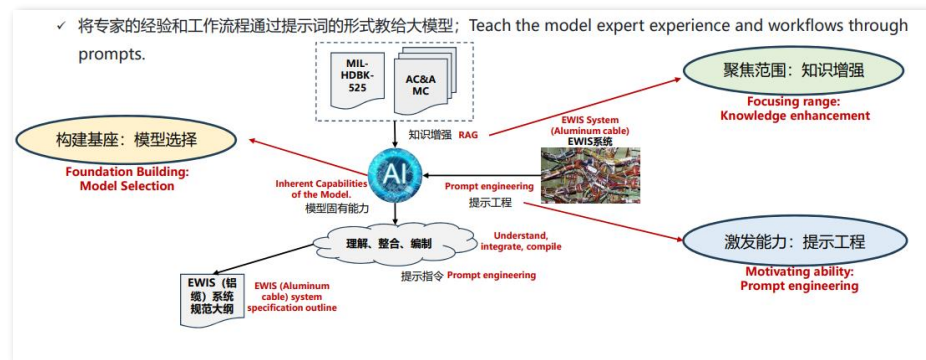
单元组合并导致数据丢失
LLM fail to recognize merged cells then lead to data loss

ПОДХОДЫ

AVIC CHINA AERO-POLYTECHNOLOGY ESTABLISHMENT



- 1) Использование LLM модели для написания стандартов (с учётом использования технологий ШАГ 1) и загрузки огромной базы международных и национальных стандартов
- 2) В рамках тестирования – 80% стандарта, соответствующу стандарту, написанному группой экспертов



ЦИФРОВАЯ ТРАНСФОРМАЦИЯ НАЦИОНАЛЬНОЙ СИСТЕМЫ СТАНДАРТИЗАЦИИ



ПОДХОДЫ РОССИЙСКОГО ИНСТИТУТА СТАНДАРТИЗАЦИИ

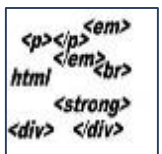
Содержание различных представлений цифрового стандарта (файлы и метаданные)



Структура и гипертекст
документа в табличной
форме



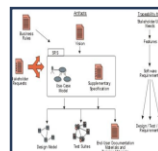
Реквизиты



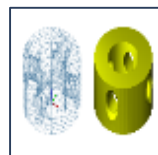
Теги разметки



Метаданные
(дополнительные атрибуты)



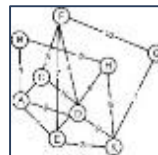
Требования



Модели



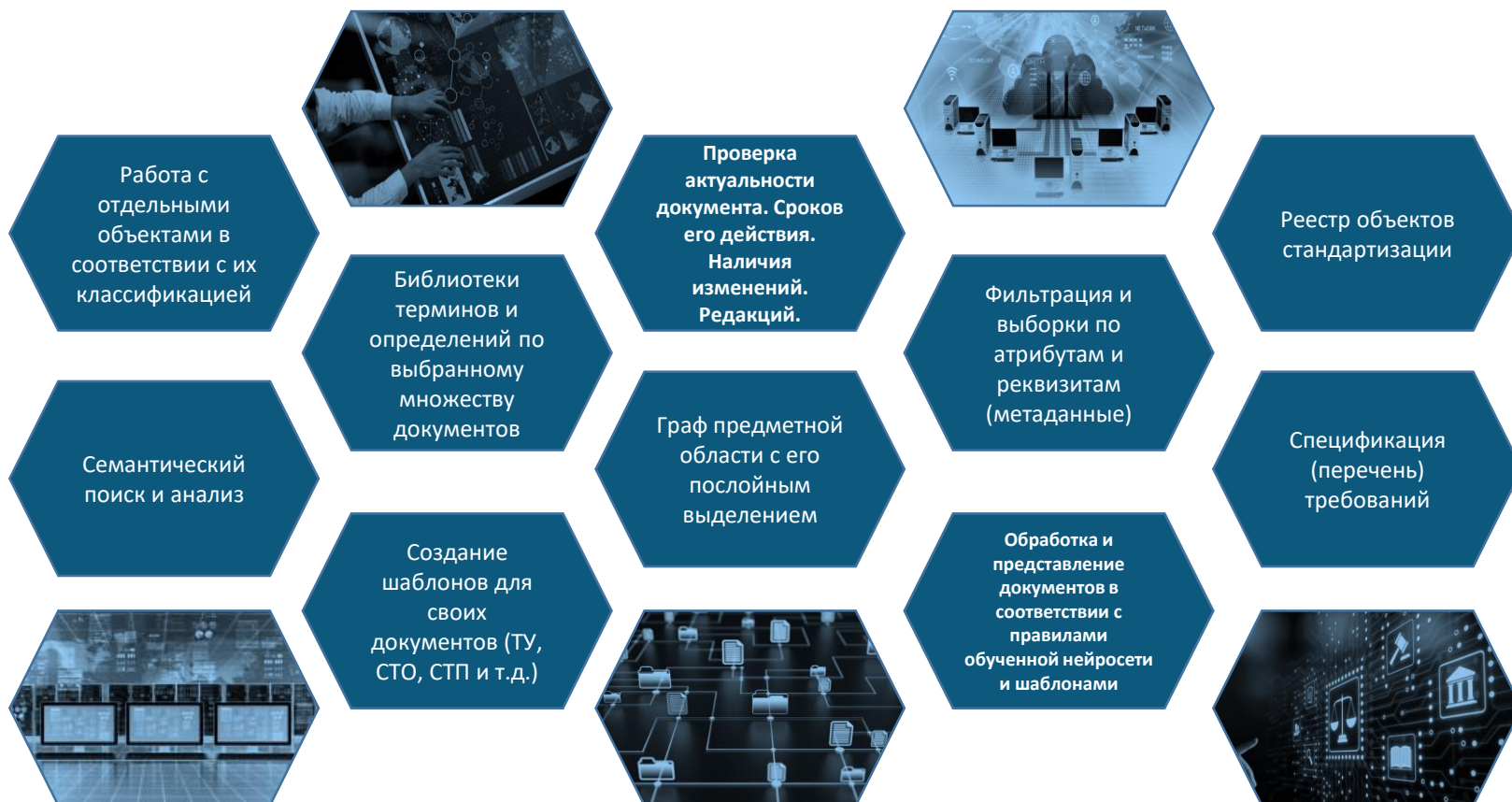
Структура (схема)
и записи БД



Граф

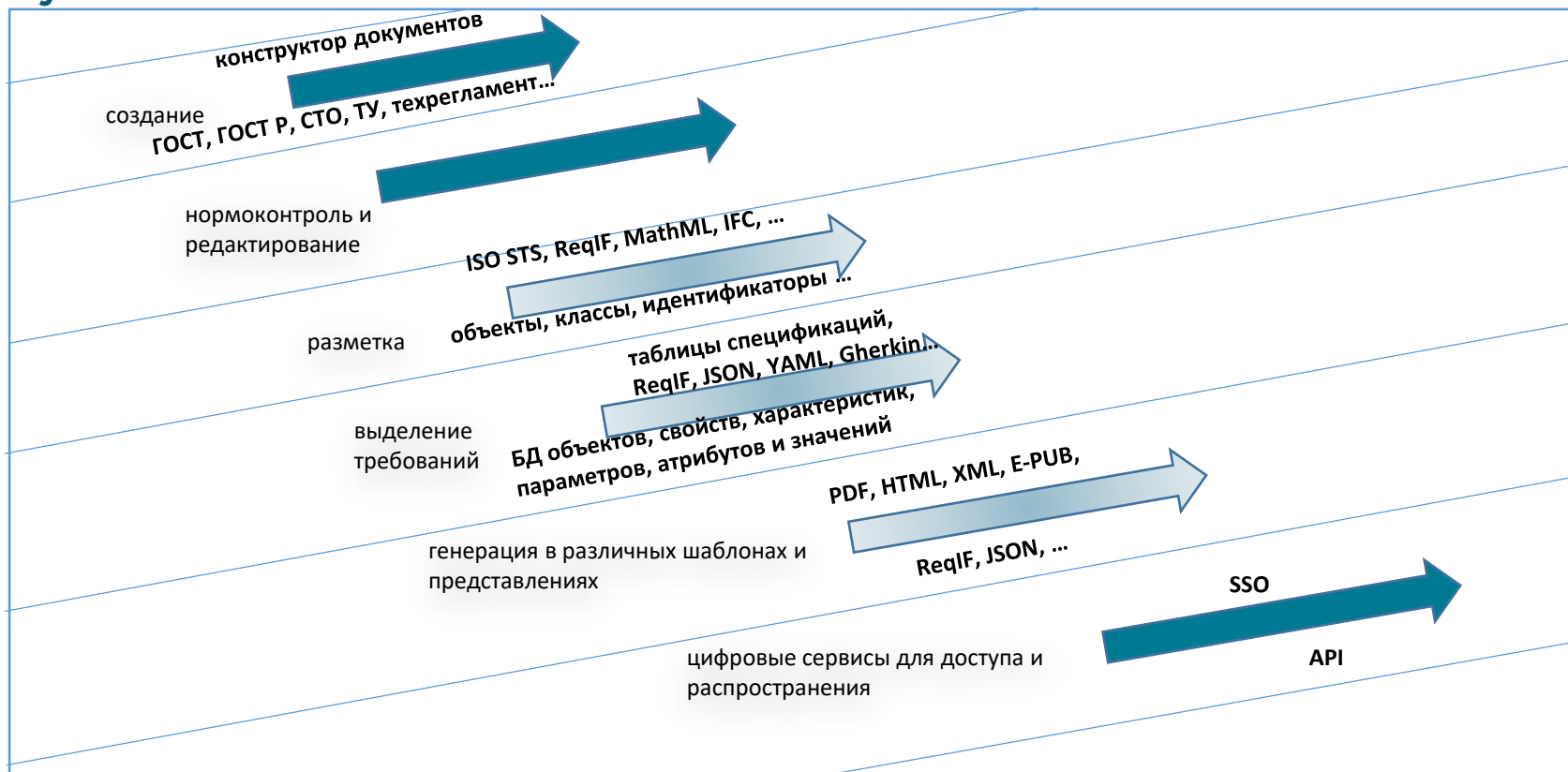
ПОДХОДЫ РОССИЙСКОГО ИНСТИТУТА СТАНДАРТИЗАЦИИ

Преимущества работы с цифровыми стандартами



ПОДХОДЫ РОССИЙСКОГО ИНСТИТУТА СТАНДАРТИЗАЦИИ

Результаты и эксперименты Российского института стандартизации. Сумма технологий



ТРЕБОВАНИЕ-ОРИЕНТИРОВАННЫЙ ПОДХОД

Требование-ориентированный стандарт в цифровом виде



ЦИФРОВОЙ ПОЛИГОН СТЕК ТЕХНОЛОГИЙ

БД Хранилища - PostgreSQL

ML – библиотеки python
(PyTorch, Threading, Translate и пр.)

OCR – Tesseract, EasyOCR,
Threading

МОДЕЛИ ОБУЧЕНЫ
на ~ 400 000
текстовых
строках,
охватывающих
~ 4500 шрифтов

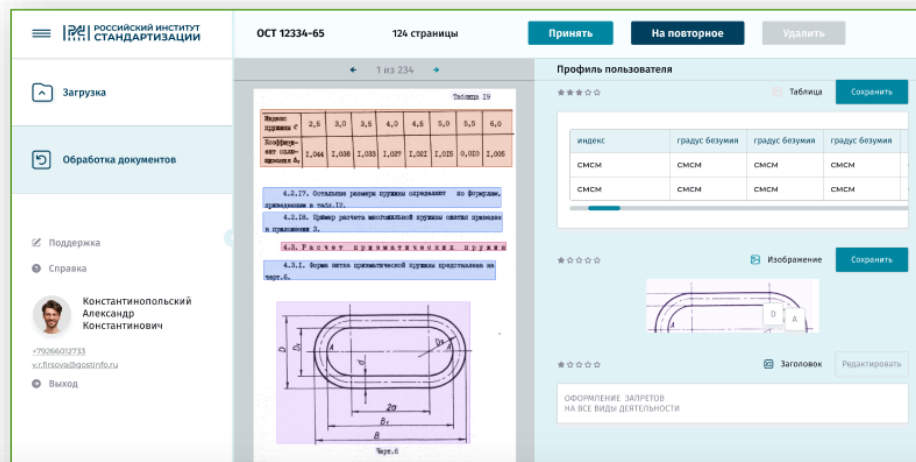
РАСПОЗНАВАНИЕ на основе Tesseract на наборах данных

Разработана
NLP модель
для классификации
текстов и предложений

Используется
Yolov7
для распознавания
таблиц и изображений

Используется
Pix2tex
для распознавания
формул

Терминология - База «Ростерм», содержащая термины из ГОСТ, ГОСТ Р, ИСО, МЭК и других документов, охватывающая более 266 000 терминов



ЦИФРОВИЗАЦИЯ КОНТЕНТА

- ✓ Формирование первой и последующих редакций проекта документа для размещения её в модуле проверки документов и прохождения всех последующих процедур работы с документом в электронном виде в соответствии с требованиями основополагающих ГОСТ.
- ✓ Формирование структуры и содержания документа с автоматической проверкой орфографии, автоматическим формированием его XML-представления, из которого формируются различные варианты электронных и печатных экземпляров.
- ✓ В качестве дополнительных опций в перспективе планируется обеспечить сервис «Конструктор документов» возможностями автоматического выделения различных сущностей, в первую очередь, терминов и определений, а также требований, правил и условий с целью формирования требование-ориентированных документов по стандартизации.
- ✓ Дополнительно планируется предусмотреть возможность расширения функционала сервиса «Конструктор документов» для решения задач создания внутренних нормативно-технических документов организаций (ТУ, СТО и т. д.).

Скриншоты интерфейса системы «Конструктор» для специалистов, демонстрирующие XML вид, Исходный pdf и OCR вид документа.

для специалистов

РОССИЙСКИЙ ИНСТИТУТ СТАНДАРТИЗАЦИИ
Информационная система «Конструктор»

ДОКУМЕНТ | ОТЧЕТЫ | ПАНЕЛЬ АДМИНИСТРАТОРА | ПОДДЕРЖКА | ПОЛЬЗОВАТЕЛЬ №10, РСТ

ДОКУМЕНТ | ИСХОДНЫЕ ФАЙЛЫ | РЕДАКТИРОВАНИЕ

РЕДАКТОР | ПОКАЗ XML | ИСХОДНИК | РЕЗУЛЬТАТ

Элементов с ошибками: 9

СОХРАНИТЬ ИЗМЕНЕНИЯ

2 / 16

2023

ГОСТ Р

Предисловие

1 РАЗРАБОТАН Акционерным обществом «Восточный научно-исследовательский углехимический институт» (АО «ВУХИИ») 2 ВНЕСЕН Техническим комитетом по стандартизации ТК 395 «Кокс и продукты коксохимии» 3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 14 декабря 2022 г. № 1483-ст

ГОСТ Р

НАЦИОНАЛЬНЫЙ СТАНДАРТ РОССИЙСКОЙ ФЕДЕРАЦИИ

XML вид

```
<title uid="1eec6819-8d23-6d60-a194-0d88e69f12e1" class="te kst">
  Предисловие
</title>
<p uid="1eec6819-8d20-6290-a194-0d88e69f12e1" class="te kst">
  1 РАЗРАБОТАН Акционерным обществом «Восточный научно-исследовательский углехимический институт» (АО «ВУХИИ»)
  </p>
  2 ВНЕСЕН Техническим комитетом по стандартизации ТК 395 «Кокс и продукты коксохимии»
  </p>
  3 УТВЕРЖДЕН И ВВЕДЕН В ДЕЙСТВИЕ Приказом Федерального агентства по техническому регулированию и метрологии от 14 декабря 2022 г. № 1483-ст
  </p>
</p>
</front>
<body uid="1eec6819-8fb2-6220-a194-0d88e69f12e1" class="te kst">
```

Исходный pdf

OCR вид

Редатор/Конструктор

РОССИЙСКИЙ ИНСТИТУТ
СТАНДАРТИЗАЦИИ



ПРИМЕНЕНИЕ ИИ ДЛЯ РАБОТЫ С ДОКУМЕНТАМИ НАЦИОНАЛЬНОЙ СИСТЕМЫ СТАНДАРТИЗАЦИИ

Формирование RAG на основе машиночитаемых стандартов

 **RAG System**
Live demo

ГЛАВНОЕ

Dashboard

ПОИСК

Поиск

Q&A

БАЗА ЗНАНИЙ

Документы

Требования

Граф связей

ОПЕРАЦИИ

Загрузка XML

Eval

Dify-тестер

СИСТЕМА

Конфиг

Настройки

Поиск команд, документов...

API OK · YandexGPT 5.1 Pro

Токен OK

ДОКУМЕНТЫ
10
в индексе

ЧАНКИ
4 634
единиц поиска

ТРЕБОВАНИЯ
990
извлечено

РЁБРА ГРАФА
622
references · replaces · harmonized

РАЗМЕР ИНДЕКСА
8.34 MB
data/processed

UPTIME
137ч 47м
openai · YandexGPT 5.1 Pro

ПОСЛЕДНЯЯ ЗАГРУЗКА
6 д назад
2026-05-05T17:13:32.485660Z

Быстрые действия

Основные пути работы с системой.

Q&A с цитированием

Вопросы по стандартам, каждый ответ с ссылкой на источник.

Гибридный поиск

Dense + BM25 поверх индекса, фильтры по документам.

Загрузка XML

Новые ГОСТы → парсер → индекс → граф.

Eval

Прогноны качества retrieval / requirements / latency.

Каталог

Стандарты, доступные в индексе.

Коксы нефтяные малосернистые. Техн...
ГОСТ 22898-2025

Изделия кладочные силикатные. Общи...
ГОСТ 379-2025

Безопасность деревообрабатывающих...
ГОСТ EN 1870-14-2016

API OK · YandexGPT 5.1 Pro

14

ПРИМЕНЕНИЕ ИИ ДЛЯ РАБОТЫ С ДОКУМЕНТАМИ НАЦИОНАЛЬНОЙ СИСТЕМЫ СТАНДАРТИЗАЦИИ

Чат-бот для ответа на вопросы

Развернутый
ответ на
вопросы
с привязкой к
стандартам

The screenshot displays the 'RAG System Live demo' interface. On the left is a sidebar menu with sections: 'ГЛАВНОЕ' (containing 'Dashboard'), 'ПОИСК' (containing 'Поиск' and 'Q&A'), 'БАЗА ЗНАНИЙ' (containing 'Документы', 'Требования', and 'Граф связей'), 'ОПЕРАЦИИ' (containing 'Загрузка XML', 'Eval', and 'Dify-тестер'), and 'СИСТЕМА' (containing 'Конфиг' and 'Настройки'). The 'Q&A' item is highlighted. The main chat area shows a search bar at the top with the text 'Поиск команд, документов...'. Below it, a status message reads: 'Каждый ответ содержит ≥1 grounded-цитату [doc::sec::elem] (ЧТЗ п. 7.6). Grounding fail виден в warnings.' The chat history shows a user question: 'какие требования к коксам нефтяным'. The chatbot's response is: 'Требования к нефтяным коксам изложены в ГОСТ 22898-2025 и включают следующие пункты: 1. Нефтяные коксы должны соответствовать стандарту и изготавливаться по утвержденной изготовителем технологии из сырья, компонентов, применявшихся при изготовлении образцов нефтяных коксов, прошедших испытания с положительными результатами [GOST_22898-2025::sec_4::p_4.1]. 2. Значения показателей нефтяных коксов должны соответствовать приведенным в таблице 1 [GOST_22898-2025::sec_4::p_4.2]. 3. Физико-химические показатели нефтяных коксов включают: • Массовая доля общей влаги, %, не более: 3,0 [GOST_22898-2025::sec_4::tbl_572e9736_row_1]. • Массовая доля летучих веществ, %, не более: варьируется в зависимости от марки кокса [GOST_22898-2025::sec_4::tbl_572e9736_row_2]. • Зольность, %, не более: варьируется в зависимости от марки кокса [GOST_22898-2025::sec_4::tbl_572e9736_row_3]. • Действительная плотность после прокалывания при температуре 1300 °C в течение 5 ч, г/см³: варьируется в зависимости от марки кокса [GOST_22898-2025::sec_4::tbl_572e9736_row_8]. • Оценка микроструктуры, балл, не менее: не нормируется, определение обязательно [GOST_22898-2025::sec_4::tbl_572e9736_row_14].'. At the bottom, there is a text input field with the placeholder 'Задайте вопрос по стандартам...' and a blue 'A' icon button. The footer of the interface shows 'API OK · YandexGPT 5.1 Pro' and a '— отправить' button.

ПРИМЕНЕНИЕ ИИ ДЛЯ РАБОТЫ С ДОКУМЕНТАМИ НАЦИОНАЛЬНОЙ СИСТЕМЫ СТАНДАРТИЗАЦИИ

Хранилище документов в виде выделенных требований и связей

Структура документа

The screenshot displays the RAG System interface with a sidebar on the left containing navigation links: Главная, Dashboard, Поиск, Q&A, База знаний, Документы, Требования, Граф связей, Загрузка XML, Eval, Dif-тестер, Система, Конфиг, and Настройки. The main content area shows the document 'ГОСТ 22898-2025' with tabs for Обзор, Структура (328), Требования (30), and Связи (134). The 'Требования' tab is active, displaying a list of requirements. A blue arrow points from the 'Структура документа' text to the 'Структура' tab. Another blue arrow points from the 'Требования' text to the 'Требования' tab. The requirements list includes items such as 'массовая доля летучих веществ для нефтяного кокса марки КЗА первого сорта' and 'выработка кубового нефтяного кокса, соответствующего марки КЗА'.

РAG System
Live demo

Документы > ГОСТ 22898-2025
ГОСТ 22898-2025
Коксы нефтяные малосернистые. Технические условия

Обзор Структура (328) **Требования (30)** Связи (134)

Поиск по тексту или связи...

3 Классификация - 1

equation $m_{\text{с.3}}$

8 Методы испытаний - 7

equation 8.3 Определение массовой доли летучих веществ > 8.3.3 Обработка результатов испытаний

$$V_A = \frac{(m_2 - m_1) \cdot 100}{m_2 - m_1} = W_A, (1)$$

equation 8.4 Определение массовой доли влаги > 8.4.3 Обработка результатов

$$X = \frac{m_1 \cdot 100}{m} (2)$$

equation 8.5 Определение действительной плотности > 8.5.2 Подготовка к проведению испытаний

допускается массовая доля летучих веществ для нефтяного кокса марки КЗА первого сорта, получаемого из нефти с содержанием парафинов более 6 % — быть не более = 9,5 %
условие: для нефтяного кокса марки КЗА первого сорта, получаемого из нефти с содержанием парафинов более 6 %

допускается массовая доля ванюды для нефтяного кокса марки КЗА первого сорта, получаемого из нефти с содержанием парафинов более 6 % — быть не более = 0,020 %
условие: для нефтяного кокса марки КЗА первого сорта, получаемого из нефти с содержанием парафинов более 6 %

допускается выработка кубового нефтяного кокса, соответствующего марки КЗА — без ограничения верхнего размера кусков
условие: кубовый нефтяной кокс соответствует марке КЗА

допускается массовая доля общей влаги — быть до = 10,0 %

допускается массовая доля серы для нефтяного кокса марки КНПС-КМ — быть не более = 0,6 %
условие: по согласованию с потребителем

должна печь — поддерживать постоянную температуру = 850 °C
условие: = 10

должна печь — обеспечивать повторный подъем указанной температуры = 3 мин
условие: не более чем за

Требования

ПРИМЕНЕНИЕ ИИ ДЛЯ РАБОТЫ С ДОКУМЕНТАМИ НАЦИОНАЛЬНОЙ СИСТЕМЫ СТАНДАРТИЗАЦИИ

Хранилище документов в виде графа связей

RAG System
Live demo

ГЛАВНОЕ

- Dashboard

ПОИСК

- Поиск
- Q&A

БАЗА ЗНАНИЙ

- Документы
- Требования
- Граф связей**

ОПЕРАЦИИ

- Загрузка XML
- Eval
- Dify-тестер

СИСТЕМА

- Конфиг
- Настройки

Поиск команд, документов...

🔍 K

API OK - YandexGPT 5.1 Pro 🟢

🔑 токен ok

Граф связей

Рёбра replaces / references / harmonized_with / ... из <std-xref> и LLM-классификации.

ТИП СВЯЗИ

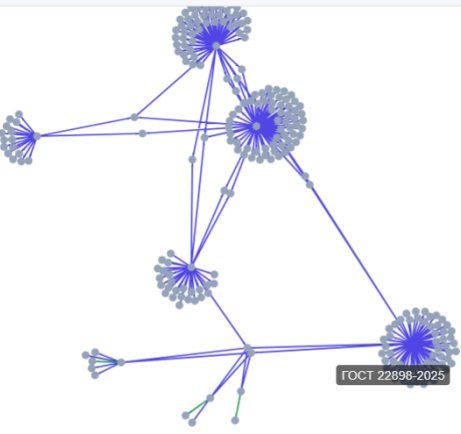
- harmonized_with
- **references**
- replaces

Сбросить

ORIGIN

- all
- xml**
- llm

272 узлов · 619 рёбер



ВЫБРАННЫЙ УЗЕЛ

ГОСТ 22898-2025
GOST_22898-2025

Открыть карточку

СОСЕДИ (133)

references

GOST_1.0

references

GOST_1.2

references

GOST_22898-78

references

GOST_12.1.007

references

GOST_12.4.010

references

GOST_12.4.011

references

GOST_12.4.068

references

GOST_12.4.103

references

GOST_12.4.252

references

GOST_12.4.310

references

GOST_17.2.3.02



РОССИЙСКИЙ ИНСТИТУТ
СТАНДАРТИЗАЦИИ

К стандартам нужен особый подход

Стандарты написаны как нормативные тексты с жёсткой структурой, поэтому обычные методы обработки часто дают сбои. Цифровизация делает требования доступнее и повторно используемыми, но только при сохранении их смысла, формы и юридической силы.



Контур проблемы цифровой трансформации стандартов на основе подхода управления требованиями с применением ИИ

01

Формальная природа

Нормативные документы соединяют формальную композицию, терминологическую точность и многоязычную среду. Для ИИ это означает необходимость работать не только с текстом, но и с юридически значимым контекстом его применения.

03

Структура и связи

Сегментация и интерпретация должны учитывать иерархию документа, кросс-ссылки и внутреннюю согласованность. Без восстановления связей стандарт превращается в набор фрагментов, утративших нормативную целостность.

02

Извлечение требований

Извлечение требований требует надёжного распознавания структурных единиц, терминов и логических связей. Ошибка на этом этапе приводит к потере обязательности, неверной атрибуции и искажению содержания нормы.

04

Семантическая эквивалентность

Перевод в цифровой вид должен сохранять не только формулировки, но и область действия, модальность и ссылки на источники. Иначе цифровая модель становится удобной, но ненадёжной.

Стандарты как иерархические и сетевые тексты

Иерархия элементов

Стандарт обычно строится как многоуровневая система: разделы, пункты, подпункты, таблицы, формулы, приложения и примечания образуют разные уровни значимости и чтения.

Функции фрагментов

Каждый структурный блок выполняет собственную функцию: одни задают обязательные требования, другие уточняют условия применения, третьи фиксируют справочные или пояснительные сведения.

Сетевые взаимосвязи

Кросс-ссылки связывают отдельные положения с другими документами и внутренними разделами. Поэтому обработка одного стандарта требует восстановления сети отсылок в пределах корпуса.

Контекст корпуса

Для корректной цифровизации важно учитывать не только локальный фрагмент, но и его место в системе нормативных документов. Иначе теряется смысл ссылок и зависимых требований.



Лексические источники ошибок извлечения

1. Многозначность терминов

Лексическая многозначность создаёт риск подмены понятий: одно и то же слово в общенаучном и нормативном употреблении может обозначать разные сущности и требования.

2. Смешение терминов

Омонимы, паронимы и редкая специализированная лексика особенно опасны для моделей, обученных на общих корпусах. Они затрудняют распознавание терминов и выделение релевантных фрагментов текста.

3. Редакционная вариативность

Вариативность редакций и отсутствие терминов в открытых корпусах приводят к нестабильным результатам извлечения. В итоге система может трактовать требования по-разному в близких версиях стандарта.

Ключевые синтаксические трудности



Номинализация

Номинализованные конструкции и отглагольные существительные делают нормативный текст плотным и трудноразборчивым. Общие синтаксические модели нередко неверно определяют главные и зависимые элементы предложения.



Порядок слов

Специфический порядок слов в стандартах меняет привычные для модели связи между членами предложения. Из-за этого нарушается правильное построение зависимостей и границ синтаксических групп.



Пунктуационные сдвиги

Нетипичная пунктуация усложняет автоматическое выделение предикативных частей и перечислений. Для стандарта важна не только грамматика, но и функциональная роль каждого знака препинания.



Смещение домена

Модели, обученные на общих текстах, часто переносят бытовые шаблоны на нормативный дискурс. В результате синтаксический анализ становится формально корректным, но содержательно ошибочным.

Семантика и прагматика нормы

01

Модальность нормы

Нормативная модальность задаёт различную степень обязательности: «должен», «следует», «может», «рекомендуется». Неправильное различение этих маркеров меняет правовой статус положения.

03

Скрытые отсылки

Имплицитные ссылки связывают положение с другими документами и разделами, но не всегда выражены явно. Для извлечения нужно восстанавливать скрытые нормативные связи по контексту.

02

Условия применения

Условные конструкции определяют область действия требования и требуют выделения антецедента и консеквента. Без этого цифровая модель не понимает, когда норма действительно применяется.

04

Адресат и действие

Прагматический уровень отвечает за круг адресатов, сроки и пределы действия нормы. Ошибка в этой зоне ведёт не к неточности формулировки, а к неверному применению требования.

Матрица рисков извлечения данных

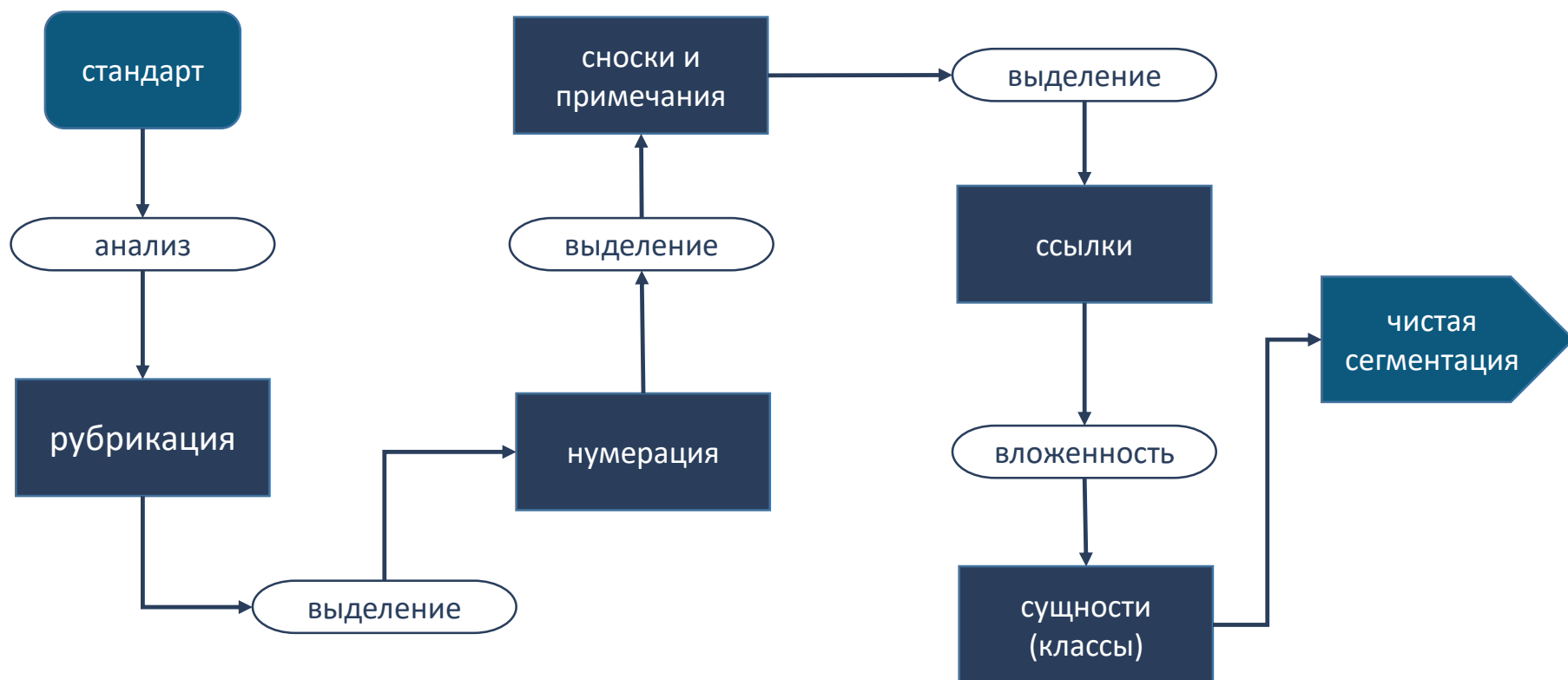
Матрица сопоставляет уровень проблемы, тип ошибки и практическое последствие для цифровой обработки стандартов.

Чем выше уровень языковой интерпретации, тем тяжелее последствия:
от терминологической путаницы до ошибочного применения нормы.

Уровень	Ошибка	Последствие
Лексика	Подмена терминов	Искажение понятий
Синтаксис	Неверный парсинг	Потеря структуры
Семантика	Сдвиг смысла	Искажение требований
Прагматика	Ошибка применения	Неверная норма

Проблемы структурной сегментации

Различия ГОСТ Р, ГОСТ, ISO, ASTM и DIN, а также вложенность пунктов и нетипичные элементы усложняют автоматическое разбиение и классификацию текста.



Стандарты как источник разнородных данных

Табличные данные

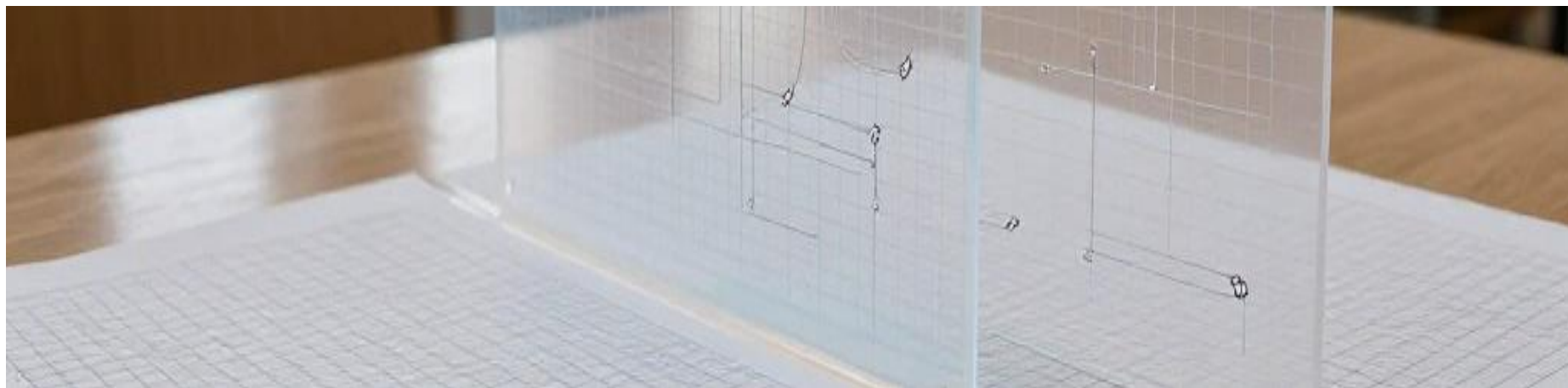
Стандарты часто содержат классификации, параметры, условные обозначения и сопоставительные данные. Для извлечения таких сведений нужны разные правила распознавания и последующей нормализации.

Графические элементы

Схемы, чертежи и диаграммы нередко несут ключевую смысловую нагрузку. Их обработка требует компьютерного зрения, распознавания условных обозначений и связывания изображения с текстовым контекстом.

Формулы и уравнения

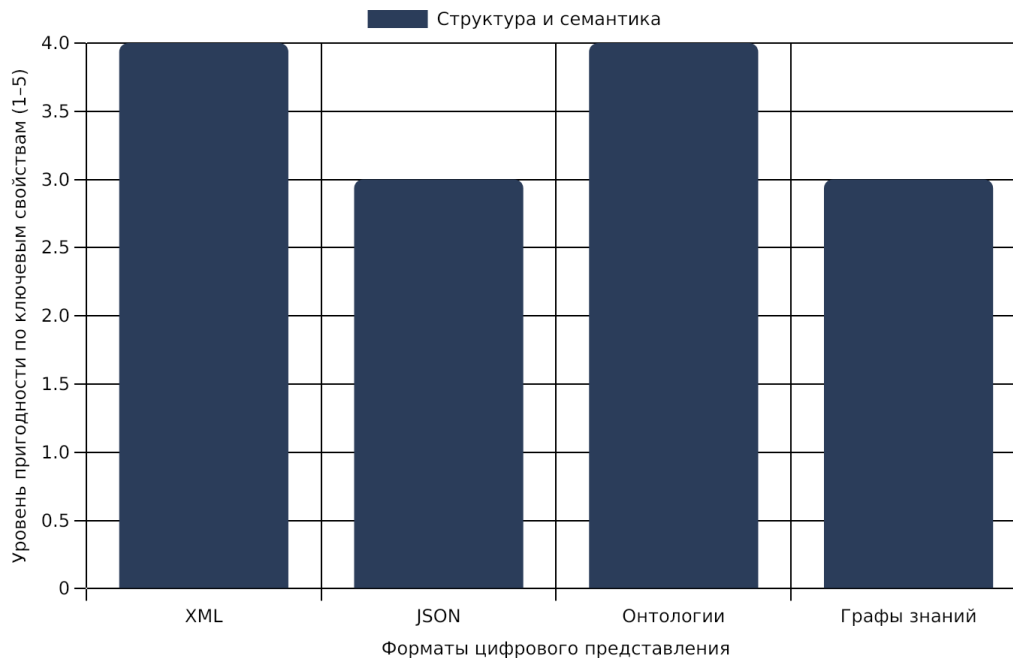
Формулы и уравнения требуют не только OCR, но и семантической интерпретации символов. Без этого цифровое представление стандарта остаётся неполным и трудно проверяемым.



Форматы цифрового представления и их ограничения

На практике выбор формата определяет не только способ хранения, но и пределы автоматической интерпретации нормативного текста.

XML лучше фиксирует структуру, а онтологии - смысловые связи; однако без валидации и согласования модели любая цифровая форма остаётся неполной.



Ключевые элементы методологического каркаса

1.

Иерархия требований

Начинать следует с разбиения нормативного текста на уровни требований: от международного или отраслевого положения к конкретной спецификации. Это создаёт основу для трассируемости и сопоставимости решений.

2.

Уникальная идентификация

Каждому требованию нужен устойчивый цифровой идентификатор, чтобы связать формулировку с источником, редакцией и статусом проверки. Без этого цифровое представление быстро теряет управляемость и доказательность.

3.

Контекст применения

В карточке требования необходимо фиксировать тип, область применения и условия действия. Эти атрибуты позволяют различать обязательные, рекомендательные и факультативные положения и корректно применять их в контексте.

4.

Связность и трассируемость

Связи между требованиями должны отражать ссылки, зависимости, дублирование и приоритеты. Такая структура помогает проследивать переход от нормы к техническому решению и выявлять противоречия между документами.

5.

Экспертная валидация

Статус валидации завершает цикл управления: требование должно быть проверено экспертом, сопоставлено с источником и признано пригодным для дальнейшей цифровой обработки и интеграции.

Что необходимо формализовать в первую очередь

Модальность нормы

Нормативные маркеры «должен», «следует», «может», «рекомендуется» нельзя сводить к одной логической форме. Для цифровой обработки требуется явное различение обязательности, допустимости и рекомендации.

Логика условий

Условные конструкции должны раскрываться как связка антецедента и консеквента. Если не выделить условие действия, система может применить требование вне предусмотренной ситуации и исказить смысл нормы.

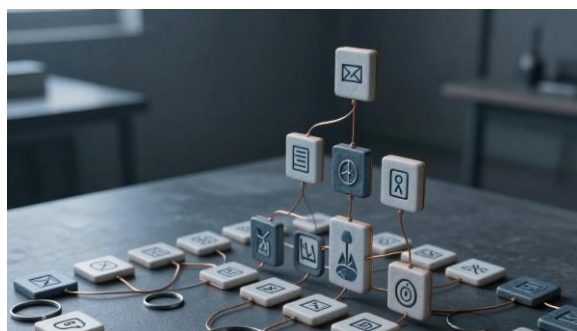
Скрытые предпосылки

Нормативный текст часто опирается на подразумеваемые знания предметной области, не выраженные прямо. Их нельзя надежно извлечь автоматически, поэтому нужны контекстные правила и экспертные комментарии.

Связи и контекст

Имплицитные ссылки на другие документы и положения требуют не только распознавания, но и проверки согласованности. Иначе цифровое представление теряет полноту и нарушает ссылочную целостность.

Три аспекта онтологического подхода



Карта понятий и связей

Онтология стандартизации начинается с карты понятий: терминов, их иерархий и ассоциаций. Когда один и тот же объект описан по-разному в нескольких документах, именно эта карта помогает увидеть расхождения и связать их в общую модель.



Извлечение с проверкой

Автоматическое извлечение онтологических структур обещает ускорение, но не снимает задачи проверки. Термины нужно сопоставить, отношения - подтвердить, а иерархии - согласовать с текстом стандарта и практикой применения.



Единое пространство без потери различий

Интеграция онтологий особенно сложна там, где одинаковые слова обозначают разные сущности. В такой ситуации цифровая модель должна учитывать вариативность терминологии, противоречивость источников и различие предметных областей.

Выводы и условия успеха



Эффективная цифровая трансформация стандартов возможна только при совместном использовании NLP, онтологий, управления требованиями и экспертной валидации. Ключ к практическому успеху — поэтапность, трассируемость и сохранение семантической точности.

БЛАГОДАРЮ ЗА ВНИМАНИЕ!

Сергей Трофимов |

советник генерального директора
Российского института стандартизации