НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Научно-учебная лаборатория учебных корпусов

РАЗРАБОТКА ПОДСИСТЕМЫ ЭКСПОРТА-ИМПОРТА КОРПУСОВ ТЕКСТОВ

Подготовил: Зайцев Константин, стажёр-исследователь

Пермь, 2020

ПОСТАНОВКА ЗАДАЧИ

Разработка хранилища корпусов текстов в гетерогенных форматах, поддерживающего возможности интеграции в сервисно-ориентированную архитектуру и импорта-экспорта корпусов в различных форматах.

ПОНЯТИЕ КОРПУСНОЙ ЛИНГВИСТИКИ

Определение

Корпусная лингвистика опирается на понятии корпуса.
Корпус – множество текстов, соответствующее определённым свойствам.
Свойства:

1. *Электронный*. Современный корпус должен храниться в электронном виде.
2. *Репрезентативный*. Корпус должен отражать изучаемый язык в полной мере.
3. *Размеченный*. Корпус должен содержать определённую разметку (синтаксическую, морфологическую, семантическую).
4. *Целеориентированный*. Каждый корпус создаётся для конкретных целей исследований.

ПОНЯТИЕ КОРПУСНОЙ ЛИНГВИСТИКИ

Корпусные менеджеры

Для работы с корпусами используются различные инструменты работы с естественным языком. В число возможностей инструментов входят следующие операции:

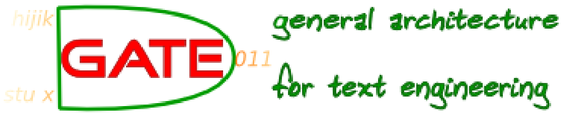
1. Поиск словоформ по определённым параметрам.
2. Отображение информации о тексте.
3. Вывод результатов поиска.
4. Получение статистических данных.
5. Сохранение конкорданса (представление искомой словоформы в контексте и со статистическими данными).

ПОНЯТИЕ КОРПУСНОЙ ЛИНГВИСТИКИ

Примеры инструментов работы над корпусами

Для представления корпусов могут использоваться следующие инструменты:

1. Sketch Engine (www.sketchengine.co.uk).
2. Gate Developer (https://gate.ac.uk/family/developer.html).
3. BRAT (https://brat.nlplab.org/).



ПРИМЕРЫ КОРПУСОВ

Примеры русскоязычных корпусов

Национальный корпус русского языка (НКРЯ)

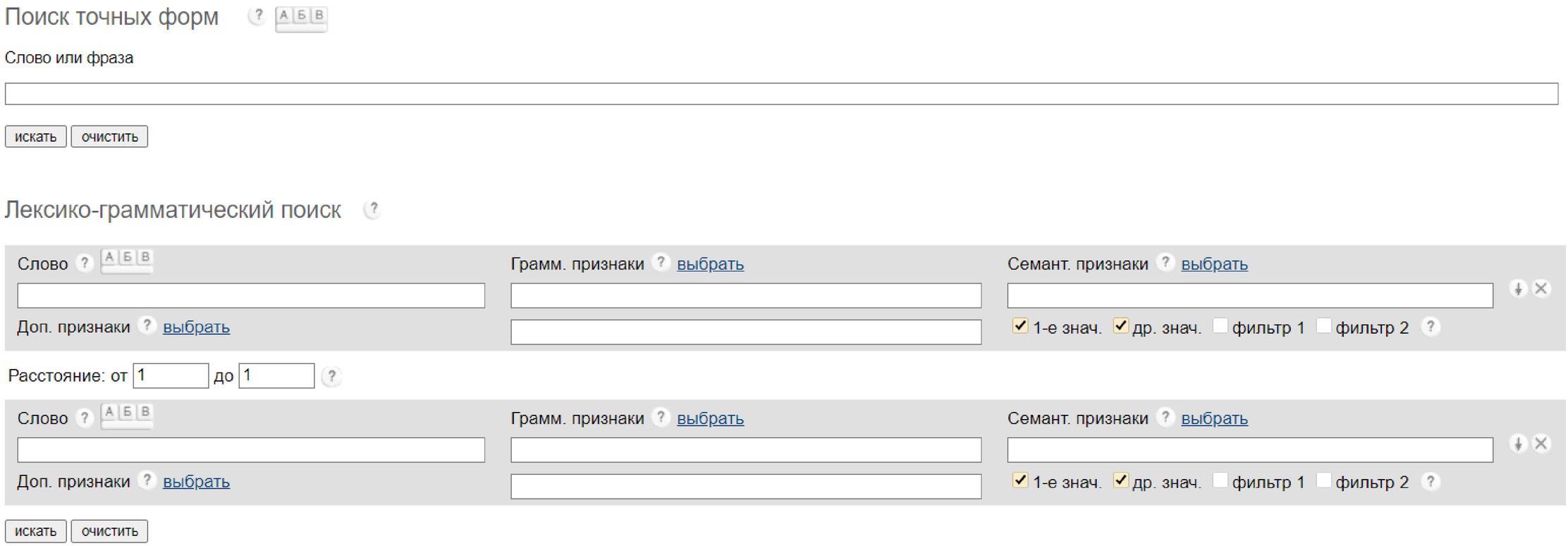
Проект поддерживается лингвистами из разных городов (Москва, Воронеж, Санкт-Петербург, Казань и т.д.) и институтов.

В Национальный корпус включены литературные тексты (с начала 18 века), переводные сочинения, поэтические тексты, тексты, представляющие нелитературные формы языка – разговорную и диалектную.

Для каждого подкорпуса имеется возможность поиска.

ПРИМЕРЫ КОРПУСОВ

Примеры русскоязычных корпусов



ПРИМЕРЫ КОРПУСОВ

Примеры русскоязычных корпусов

Объем всего корпуса: 124 565 документов, 321 712 061 слово.

"корпус"

Найдено 3 097 документов, 8 446 вхождений.

Распределение по годам Статистика

Поискать в других корпусах: акцентологическом, газетном, диалектном, мультимедийном, обучающем, параллельном, поэтическом, устном.

Страницы: **1** 2 3 4 5 6 7 8 9 10 11 следующая страница

1. **И. А. Барков. Автоматический синтез структурного описания конструкции // «Информационные технологии», 2004** [омонимия снята] **Все примеры**

   В приведённом примере описания у конструкторского понятия "сборочная единица" раздел Состав присутствует, а у обозначающих детали понятий "**корпус**" "втулка" "вал" раздел Состав пуст. [И. А. Барков. Автоматический синтез структурного описания конструкции // «Информационные технологии», 2004] [омонимия снята] ←…→
   В примере понятия "сборочная_ единица" "**Корпус**" "Втулка" "Вал" отражают конструкторские представления о свойствах изделия, а понятия "Размерная_ цепь" "Размер" "Оборудование" — представления о том же изделии инженера-технолога. [И. А. Барков. Автоматический синтез структурного описания конструкции // «Информационные технологии», 2004] [омонимия снята] ←…→

2. **Михаил Кузьминский. Водопады Supermicro // «Computerworld», 2004** [омонимия снята] **Все примеры**

   Корпус сервера-лезвия с формфактором 7 U и блоками питания мощностью 4, 5 либо 3 кВт может включать в себя и модули коммутаторов, в том числе InfiniBand. [Михаил Кузьминский. Водопады Supermicro // «Computerworld», 2004] [омонимия снята] ←…→

ПРИМЕРЫ КОРПУСОВ

Примеры русскоязычных корпусов

Открытый корпус русского языка (OpenCorpora)

OpenCorpora – это проект по созданию размеченного корпуса силами сообщества.
Размеченные корпуса находятся в открытом доступе, каждый может скачать их и поработать с ними.
В проекте может поучаствовать любой заинтересованный человек, достаточно зарегистрироваться и выбрать задание.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

British National Corpus (BNC)

Британский национальный корпус – собрание текстов на английском, включающий письменную и устную речь.

Корпус создан при помощи издательства Оксфордского университета.

Письменная часть корпуса представлена большим количеством жанров: новостными газетами, журналами, научными публикациями, эссе и многое другое.

Разговорный подкорпус содержит транскрибированные неформальные разговоры, формальные бизнес встречи или государственные мероприятия, передачи на радио и т.д.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

Open American National Corpus (OANC)

Американский национальный корпус – корпус английского языка, в который входят письменные и устные источники – сообщения на почте, твиты, текст с веб-страниц.
Можно стать участником в формировании и аннотации корпуса.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

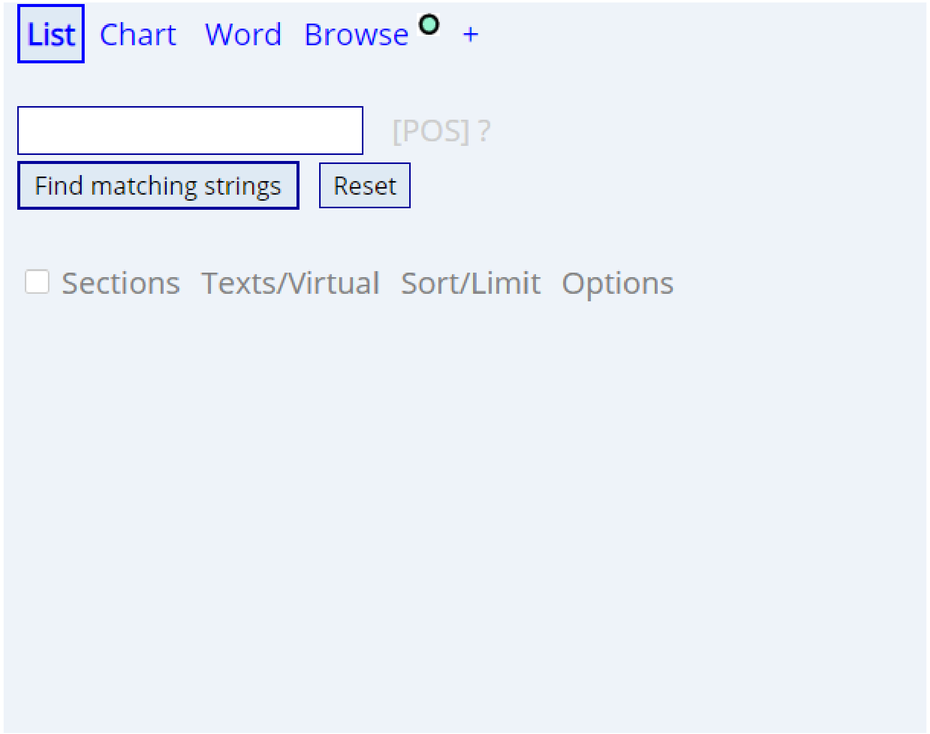
Corpus of American English (COCA)

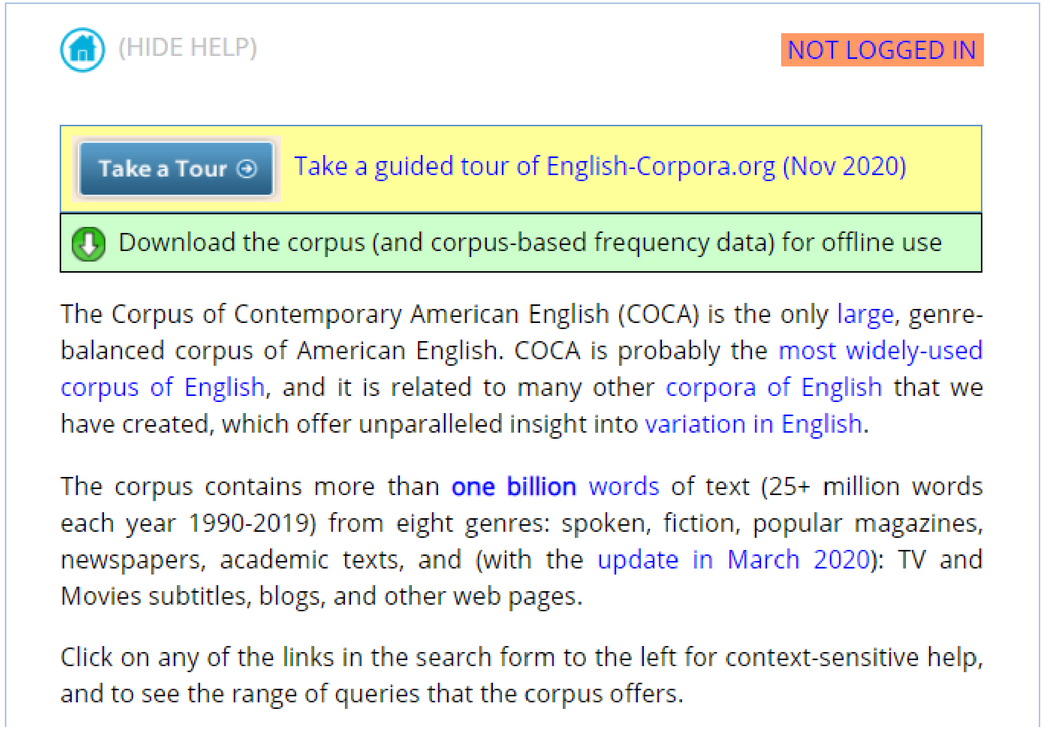
Корпус содержит огромное количество текстов (суммарно больше 25 миллионов слов с 1999 по 2019 гг.) на английском языке по разным жанрам: разговорный, художественный, популярные журналы, новостные газеты, научные тексты и, с недавнего времени, телевидение, субтитры, блоги и веб-страницы.

Имеется возможность проводить поиск по корпусу таким же образом, как в национальном корпусе русского языка.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов





ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

FIND SAMPLE: 100 200 500 1000
PAGE: << < 1 / 41 > >>

CLICK FOR MORE CONTEXT [?] SHOW DUPLICATES

1	2012	BLOG	crawfishboxes.com	A	B	C	acquisition Oberholtzer strangely enough upped his strikeout rate from 6.6 to 9.2 after joining the **Corpus** Christi hooks, his ERA also jumped to 5.27 in 27.
2	2012	BLOG	crawfishboxes.com	A	B	C	rate (SO/9) and a 2.2 walks per nine rate (BB/9) between **Corpus** Christi and Oklahoma City. His ERA between the two levels was a less than
3	2012	BLOG	crawfishboxes.com	A	B	C	At Oklahoma he posted an average Game Score of 48.47 in 15 starts. At **Corpus** Christi he posted a Game Score of 50.62 in 13 starts. # The thing
4	2012	BLOG	...rspains.blogspot.com	A	B	C	headquarters, tried by a military court 6-7 May, denied a writ of habeas **corpus**, and sentenced to 2 years' confinement in a military prison. Following a
5	2012	BLOG	...rspains.blogspot.com	A	B	C	Parte Vallandigham, which decreed that the Court could not issue a writ of habeas **corpus** in a military case, and for a Democratic campaign slogan he crea
6	2012	BLOG	...review.wordpress.com	A	B	C	Illias Panagiotaros (an MP with the Golden Dawn who participated in the attacks outside **Corpus** Cristi last week) admitted to Newsnight's Paul Mason last '
7	2012	BLOG	reason.com	A	B	C	, I object more to the actions Lincoln took during the war (suspending habeas **corpus**, the draft, etc) than the war itself. And I'm in
8	2012	BLOG	reason.com	A	B	C	less casualties. # A hero who invaded a sovereign nation, illegally suspended habeas **corpus** and imprisoned thousands of political enemies, tacitly sanctio
9	2012	BLOG	chronicle.com	A	B	C	BYU had previously created his own POS-tagged version of Google Ngrams as part of his **corpus** collection, he only had access to the publicly available data
10	2012	BLOG	chronicle.com	A	B	C	threshold of 40 occurrences for inclusion) and thus wasn't able to parse the **corpus** in a systematic fashion. " # (The answer is putting only after he
11	2012	BLOG	laboratorium.net	A	B	C	remind people that not everyone is pleased that Google did not come by their complete **corpus** by totally honest means, in other words the foundations o
12	2012	BLOG	laboratorium.net	A	B	C	in-copyright works as currency to get a better deal for their subscription to the Google **corpus**. The University of Wisconsin is no better than a school yard b
13	2012	BLOG	laboratorium.net	A	B	C	# And then there is the question of who gets access to the complete Google **corpus** (which contains works from the world over) by subscriptions -- just Am
14	2012	BLOG	laboratorium.net	A	B	C	At the D is for Digitize Symposium?) the point that Google owns the **corpus** and could if it wanted to sell it... " even to China ".
16	2012	BLOG	laboratorium.net	A	B	C	his promises. " The settlement " gives Google the right to sell the **corpus**, " she said. " Suppose they sell it to Rupert Murdoch... Wal-Mart
17	2012	BLOG	laboratorium.net	A	B	C	to engage in lawful large-scale book digitization projects, thereby denying the public the full **corpus** of 20th century books and, perhaps, unknown benefits

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

Czech National Corpus (CNC)

Корпус состоит из собраний текстов на чешском языке. В их число входят современный чешский, разговорный язык, корпус исторических текстов и параллельный корпус, содержащий перевод на более 30 других языков.

Корпус доступен для всех, кто заинтересован в его исследовании, в нём также доступна возможность поиска интересующих словоформ.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

Czech National Corpus

Search engine demonstration:

Search for Czech words:

Search

Country:

Czech Republic

CLARIN Centre:

Charles University

Description

A Corpus is a collection of texts in electronic form used for linguistic research, using provided with digital tools to allow searching, analysis and research. Users can use these tools to find words and collocations in their original contexts, and determine their frequency in the corpus. **The Czech National Corpus** (CNC) is an academic project focusing on building a large electronic corpus of mainly written Czech. The Institute of the Czech National Corpus (ICNC), Faculty of Arts, Charles University in Prague oversees the development of the CNC, including its use in teaching, and advancing the field of the corpus linguistics.

ПРИМЕРЫ КОРПУСОВ

Примеры иностранных корпусов

Hits: **94** | i.p.m.: **0.77** (related to the whole corpus) | ARF: **44.19** | Result is shuffled

1 / 3

Line selection: simple

K znamená Kennedy	žádost odmítneme , musíme mít prezidentův podpis ke zrušení Habeas	Corpus	pro tento případ . Francisi , jsi ochoten podepsat mi
Vynálezci z Guslaru	mu dáme peníze , chytíme ho in flagranti , s	corpus	delicti v ruce . " Ráno strávili velice pilnou prací
Vzkříšení Olympie	. Vázových maleb se však zachovalo takové množství , že	Corpus	vasorum antiquorum , který je zachycuje , má od roku
Sport magazín, č. 12/2008	slovní spojení " Jagr and Granger " . Baseballová pálka	CORPUS	DELICTI Historie : Kolem roku 1850 , v počátcích baseballu
Lidové noviny, 14. 2. 2006	stabilním stavu . Oznámila to včera mluvčí nemocnice v texaském	Corpus	Christi , kam byl osmasedmdesátiletý Harry Whittington po incidentu v
Chladnokrevně	zjištěné skutečnosti opravňovaly , napadl platnost rozsudku podle zásady habes	corpus	u kansaského Nejvyššího soudu , který ho nedávno potvrdil .
Svět středověkých gest	Zatím bylo v hlavním ohnisku pozornosti věřících Tělo Páně ,	Corpus	Christi , jemuž patřila i zvláštní zbožnost . Vznikla v
Moderní řízení, č. 10/2006	útvaru , " říká Ing. Radek Bělina , konzultant společnosti	Corpus	Solutions a. s. , a autor tohoto článku a pokračuje
Živě.cz, 6. 1. 2005	zástupci lokálních partnerů ze společností Barracudanetworks , S &T a	Corpus	Solution , kteří ve svých příspěvcích využijí zkušeností se specifickými
Hold Katalánsku a Ohlédnutí za Španělskou válkou	Ve Španělsku neexistuje - alespoň se nepraktikuje - žádný habeas	corpus	a ve vězení vás mohou držet nepřetržitě několik měsíců ,
Jedinec: Gen - prostředí - vývoj	se v důsledku prudkého snížení krevní hladiny lh aktivita buněk	corpus	luteum a kolem 12 . dne po ovulaci ustane .
Encyklopedie mystiky II.	, v tomto textu popsáno není ; jiný text z	Corpus	Hermeticum je označí za " božské mlčení " . Připomeňme
Dům z úsvitu	šlo o hada . Že prej zabil pitomýho hada !	Corpus	delicti , drž se , vyhrožoval , že se promění
Silnější než slova	je to tak . Jak potom vysvětlíte tu práci v	Corpusu	? " " Dělal jste tam totéž . EastPark je

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Стандарты разметки

В ходе исследования разметок были обнаружены свои стандарты разметок и форматы представления для различных корпусов.

1. Стандарт BNC. Присутствие метаинформации, текст разбивается на предложения, предложения – на слова с частеречной разметкой.
2. Формат базы данных. Корпус разбивается на текстовые файлы. Для текста, словаря и метаинформации созданы отдельные таблицы.
3. Стандарт GATE Developer. Текст разбивается на три секции. Первая - метаинформация, текст с узлами, третья – аннотация узла (слова).

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Стандарты разметки

4. Стандарт НКРЯ. Используется HTML-разметка, текст разбивается на абзацы, которые имеют определяющий их атрибут. У каждой лексической единицы есть свой тег, который их обозначает.
5. Стандарт OpenCorpora. Присутствует метаинформация, текст разбивается на параграфы, параграфы – на предложения, предложения – на токены. Токены описываются специальными атрибутами.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов

Исходя из просмотренных стандартов и форматов, можно выделить следующие форматы хранения корпусов:

1. Формат базы данных.
2. Формат плоских файлов.
3. Размеченный формат (XML, HTML, SGML).

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов

Формат базы данных и плоских файлов

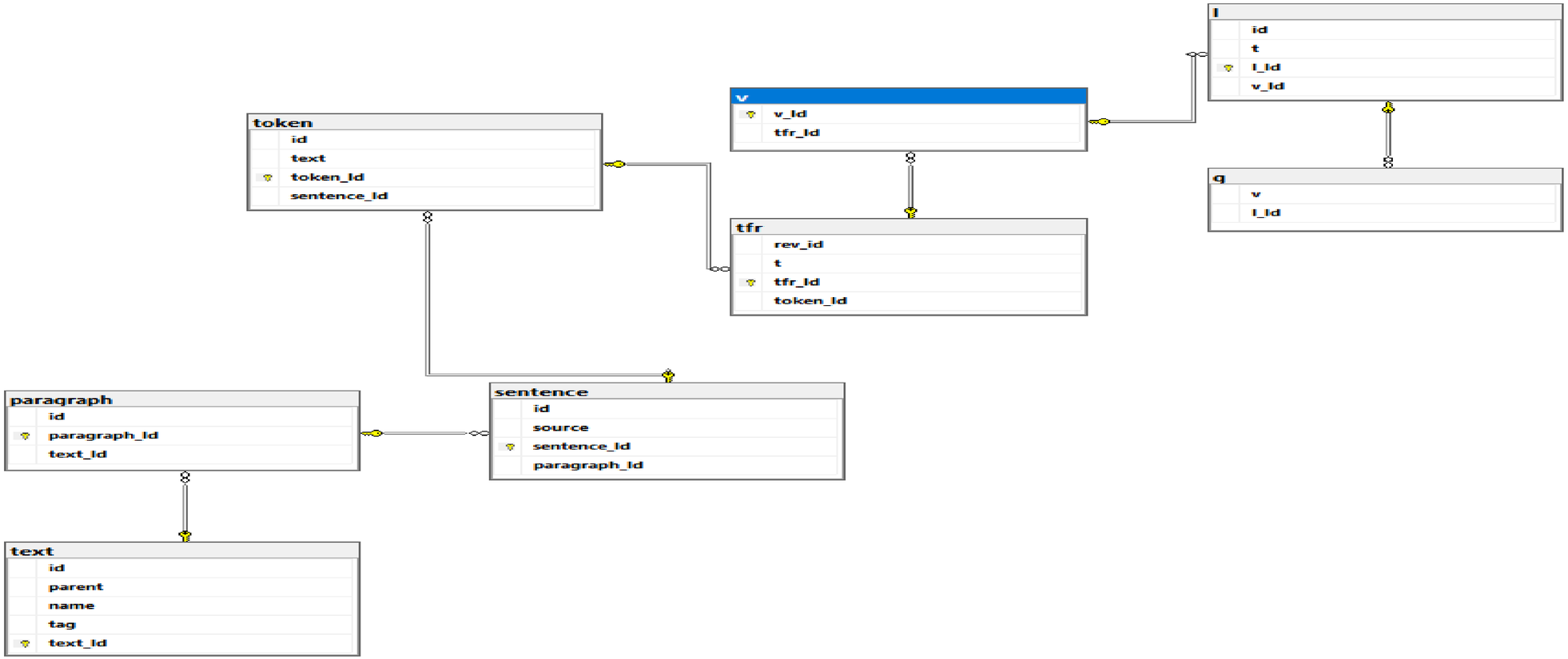
Оба формата довольно похожи, поскольку они обычно представляются в виде таблиц или файлов.

Каждый атрибут, токен, единица разметки корпуса могут быть представлены в виде отдельной таблицы. Между таблицами/файлами существует определённая связь, определяющая структуру корпуса.

В формате баз данных удобно находить слова при помощи SQL-запросов.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов



СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов

```
1 id  parent  name    tag text_Id corpus_Id
2 1   0   "Частный корреспондент" url:http://www.chaskor.ru   0   1
3 2   1   00021 Школа злословия   Год:2008    1   1
```

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов

Размеченный формат

Обычно формат представлен в виде XML-файлов, но есть варианты HTML и SGML-разметки.

Слова, предложения, метаинформация заключается в тэги. У тэгов могут быть атрибуты, которые более подробно описывают заключенную в них информацию. Например, для слов через атрибут может передаваться часть речи, падеж и прочие категориальные признаки.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Форматы хранения корпусов

```
<text id="int" parent="int" name="string">
    <tags>
      <tag>string</tag>
    </tags>
    <paragraphs>
      <paragraph id="int">
        <sentence id="int">
          <source>string</source>
          <tokens>
            <token id="int" text="string">
              <tfr rev_id="id" t="string">
                <v>
                  <l id="int" t="string">
                    <g v="string"/>
                  </l>
                </v>
              </tfr>
            </token>
          </tokens>
        </sentence>
      </paragraph>
    </paragraphs>
</text>
```

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Трансформация размеченных и плоских файлов, экспорт и импорт

Суть экспорта или импорта – преобразование разметки в текст, базу данных или другой стандарт.

В целом трансформацию можно разделить на два класса, которые соответствуют определённым форматам:

1. XSL-преобразование.
2. XSD-схема.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Трансформация размеченных и плоских файлов, экспорт и импорт

XSL-преобразование

XSLT – преобразование файла одного формата или стандарта в другой при помощи xsl-файла, в котором прописываются правила трансформации.

Для xslt-файлов определён собственный синтаксис, в котором прописываются различные шаблоны входного файла.

На выходе получается файл или несколько файлов с той разметкой, которая определена правилом.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Трансформация размеченных и плоских файлов, экспорт и импорт

```
<xsl:stylesheet version="1.0"
xmlns:xsl="http://www.w3.org/1999/XSL/Transform">

    <xsl:output method="text" indent="yes"/>
    <xsl:strip-space elements="*"/>
    <xsl:template match="/">
      <xsl:text>id&#009;text</xsl:text>
      <xsl:value-of select="'&#13;'"/>
      <xsl:apply-templates select="annotation/text/paragraphs/paragraph/sentence/tokens" />
   </xsl:template>
    <xsl:template match="token">
        <xsl:value-of select="concat(@id, '&#009;', @text, '&#13;')"/>
    </xsl:template>

</xsl:stylesheet>
```

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Трансформация размеченных и плоских файлов, экспорт и импорт

XSD-схема

Для трансформации файла с XML-разметкой можно воспользоваться XSD-схемой.

Она описывает структуру XML-файла, его тэги, атрибуты, типы данных и т.д.

Соответственно эта структура должна иметь такую же структуру, какая присутствует у базы данных.

Через комплексные типы образовываются связи между таблицами.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Трансформация размеченных и плоских файлов, экспорт и импорт

```
<xs:schema xmlns:xs="http://www.w3.org/2001/XMLSchema">
    <xs:element name="text">
        <xs:complexType>
            <xs:sequence>
                <xs:element name="tag" type="xs:string"/>
                <xs:element name="paragraph">
                    <xs:complexType>
                        <xs:sequence>
                            <xs:element name="sentence">
                                <xs:complexType>
                                    <xs:sequence>
                                        <xs:element name="source" type="xs:string"/>
                                        <xs:element name="token">
```

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

ETL-процесс

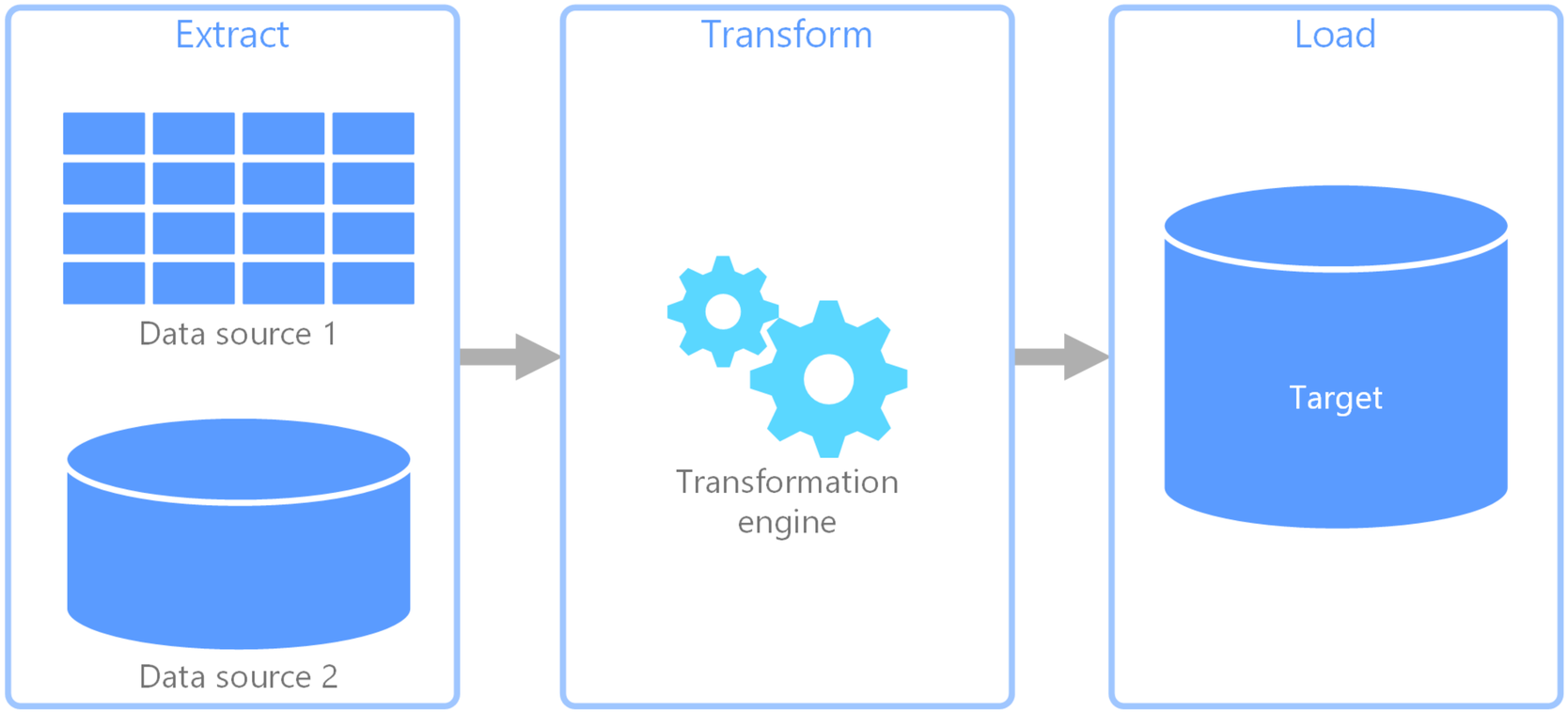
ETL - инструменты позволяют экспортировать или импортировать источники данных в назначения данных. Например, можно импортировать данные в excel-файле в базу данных.

ETL расшифровывается как Extract, Transform, Load. Подход можно описать следующим процессом:

1. Извлечение данных из источников.
2. Обработка, очистка данных, преобразование, интеграция.
3. Загрузка данных в назначение данных.

СУЩЕСТВУЮЩИЕ ПОДХОДЫ

ETL-процесс



СУЩЕСТВУЮЩИЕ ПОДХОДЫ

Примеры ETL-инструментов

ETL-процессы поддерживаются следующими инструментами:

1. SQL Server Integration Services (SSIS).
2. Oracle Data Integration.
3. GATE Developer.

ПРЕДЛАГАЕМЫЙ ПОДХОД

Основные моменты

Суть – объединение описанных подходов в решаемой проблеме.

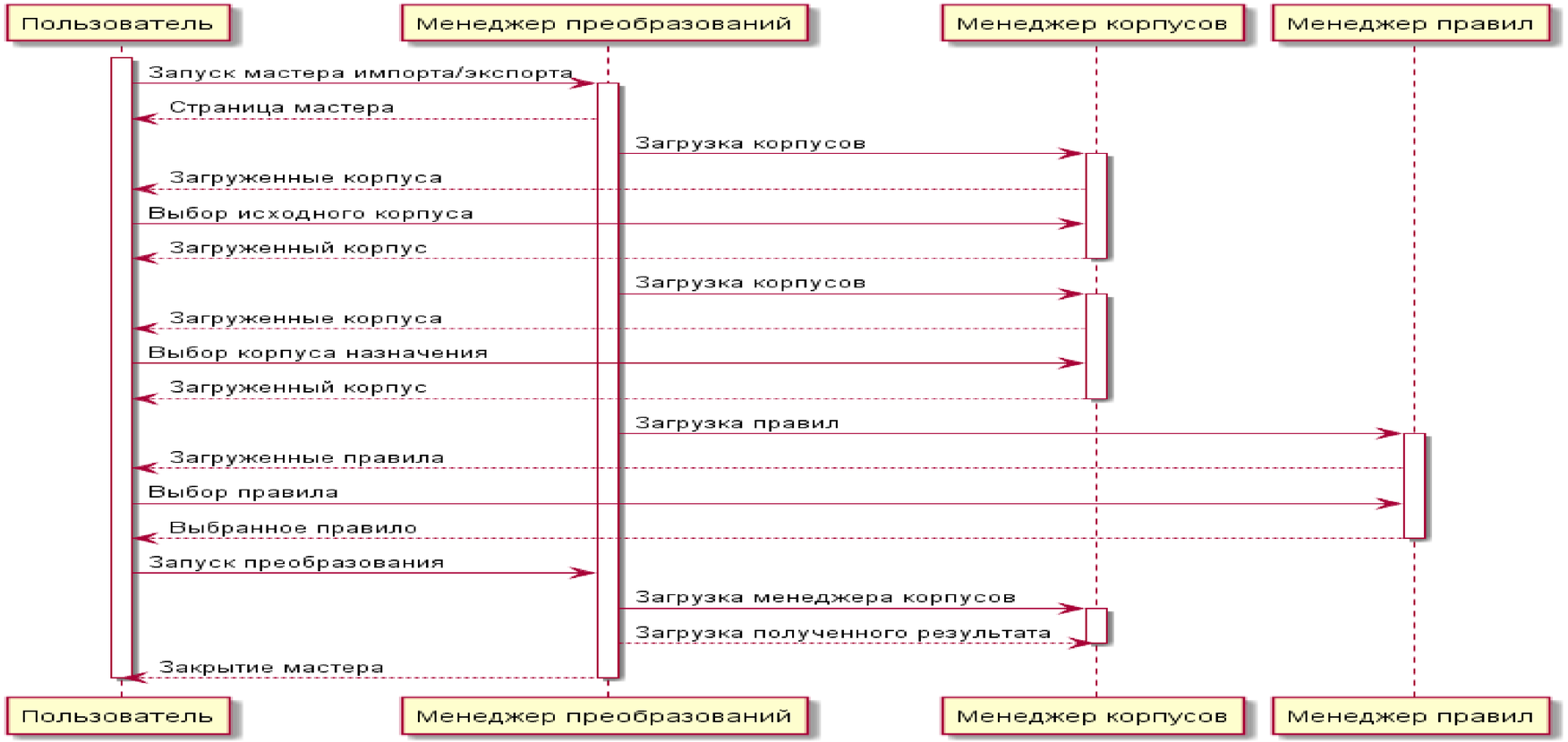
Для базы данных была разработана собственная схема, базирующаяся на стандарте OpenCorora. На нём же основывается XML-формат.

На основе ETL подхода разработан основной процесс импорта и экспорта корпусов.

Для каждой части процесса написан свой модуль.

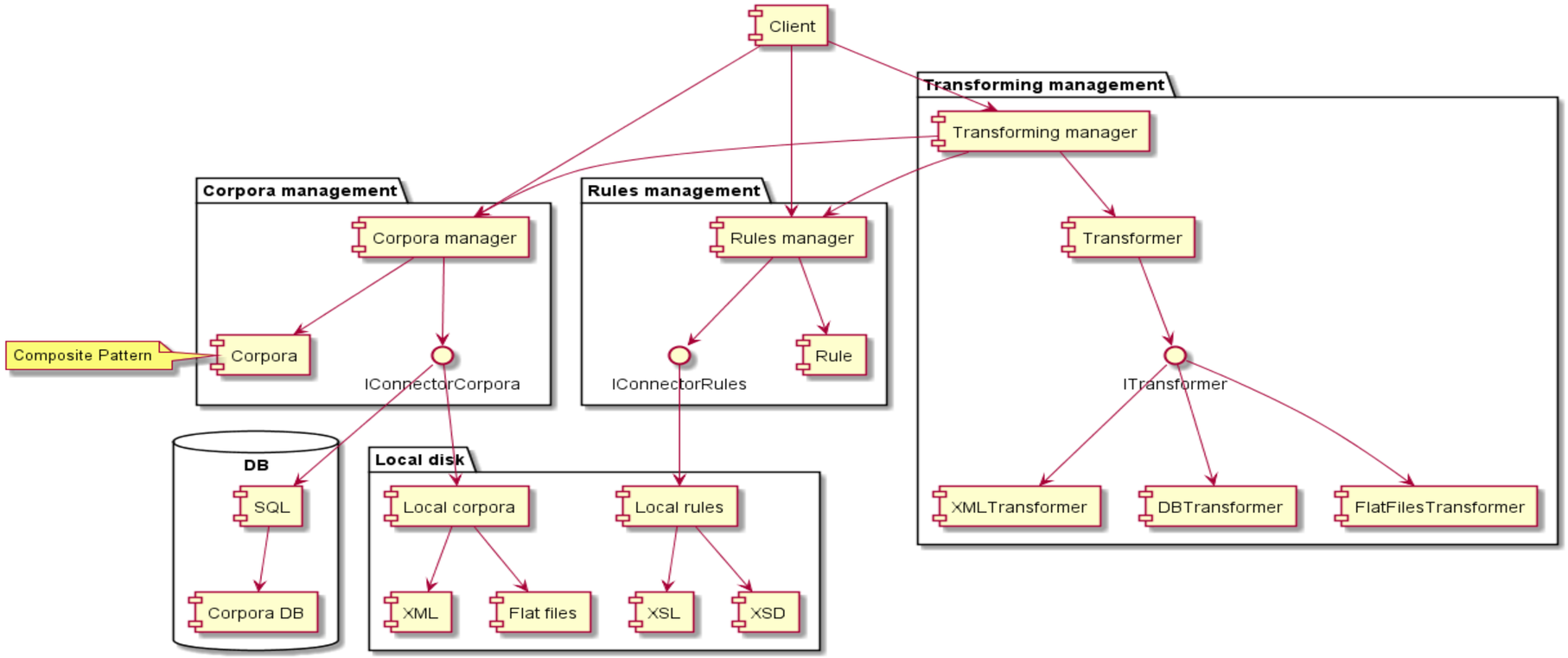
ПРЕДЛАГАЕМЫЙ ПОДХОД

Процесс экспорта и импорта



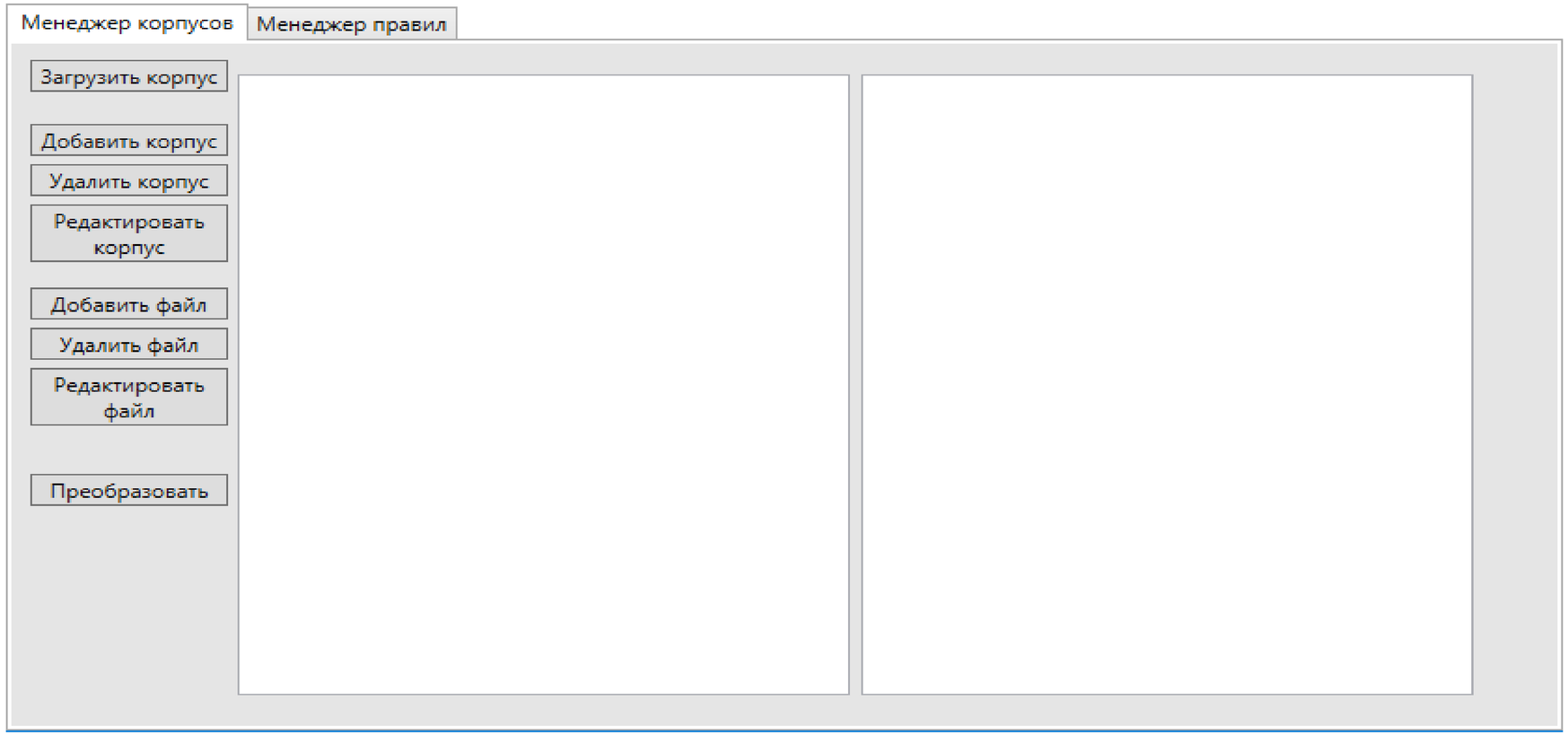
ПРЕДЛАГАЕМЫЙ ПОДХОД

Архитектура системы



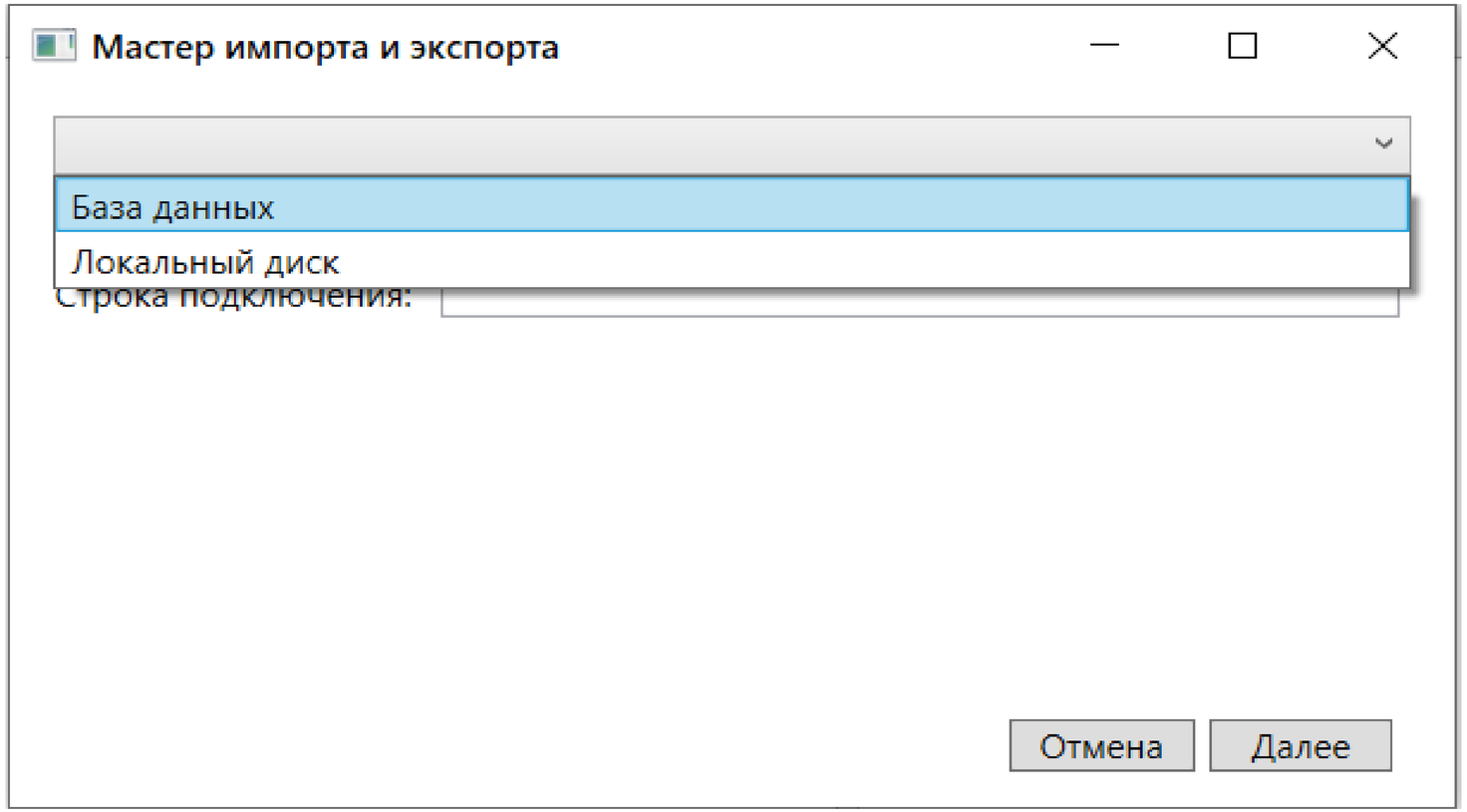
ПРЕДЛАГАЕМЫЙ ПОДХОД

Интерфейс системы. Менеджер корпусов



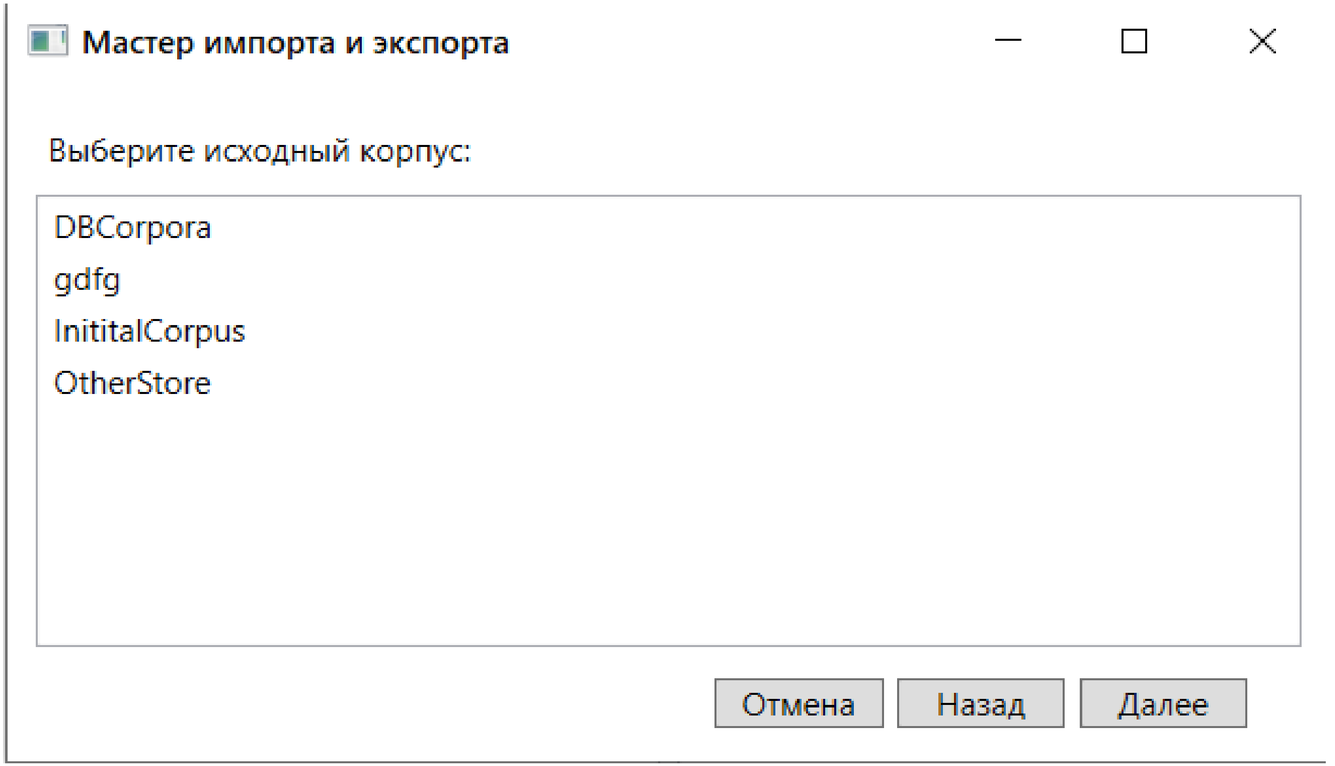
ПРЕДЛАГАЕМЫЙ ПОДХОД

Интерфейс системы. Мастер импорта и экспорта. Выбор типа хранилища



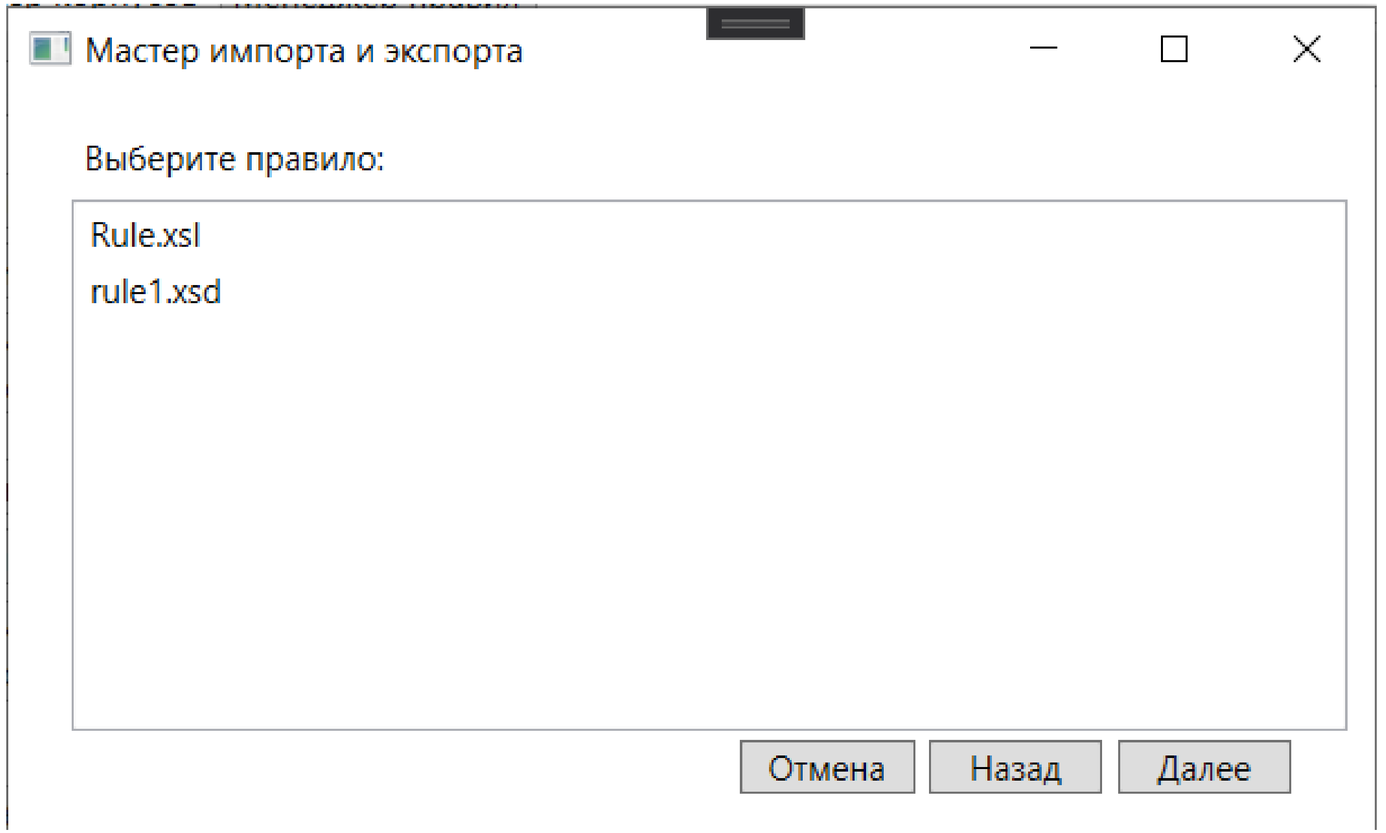
ПРЕДЛАГАЕМЫЙ ПОДХОД

Интерфейс системы. Мастер импорта и экспорта. Выбор исходного корпуса



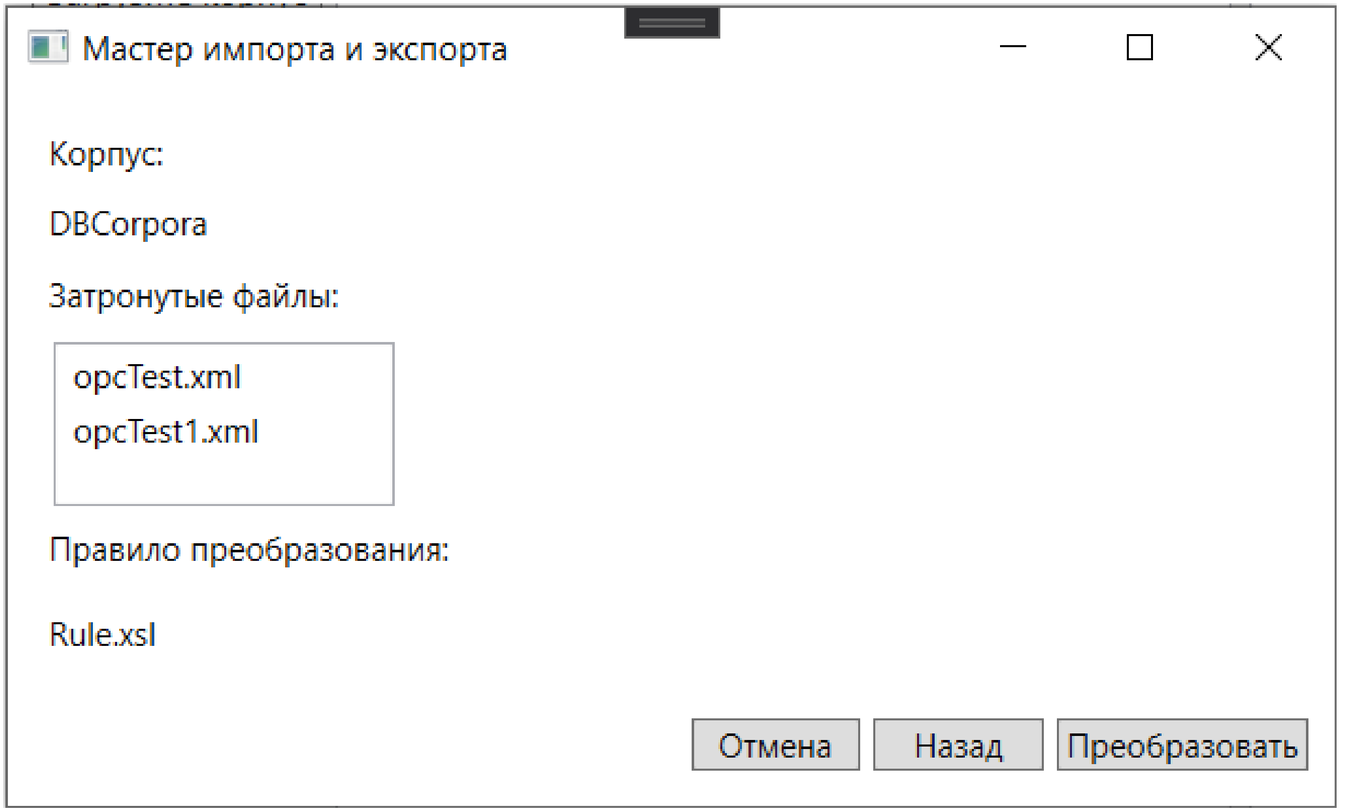
ПРЕДЛАГАЕМЫЙ ПОДХОД

Интерфейс системы. Мастер импорта и экспорта. Выбор правила



ПРЕДЛАГАЕМЫЙ ПОДХОД

Интерфейс системы. Мастер импорта и экспорта. Последняя стадия процесса преобразования



ЧТО ДАЛЬШЕ

Проблемы и будущие наработки

Проблемы:

1. Импорт в базу данных работает довольно долго.
2. Каждый файл означает текст корпуса. Некоторые системы аннотирования хранят файлы с метаданными (например, BRAT).

Что планируется:

Описанная система станет частью системы управления хранилищем корпусов на основе использования NoSQL базы данных.

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

Спасибо за внимание!