НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

НУЛ учебных корпусов

Автоматизация процессов импорта-экспорта в системах направленных на обработку корпусов текстов

Салахов Роман Данилович

Пермь, 2020

План доклада

1. Устройство корпусов
2. Распространенные системы работы с корпусами
3. ETL системы

Устройство корпусов

Корпус – совокупность текстов подобранная по определенному правилу.

Части корпуса:

- Текст
- Аннотация (разметка)

Разметка корпусов

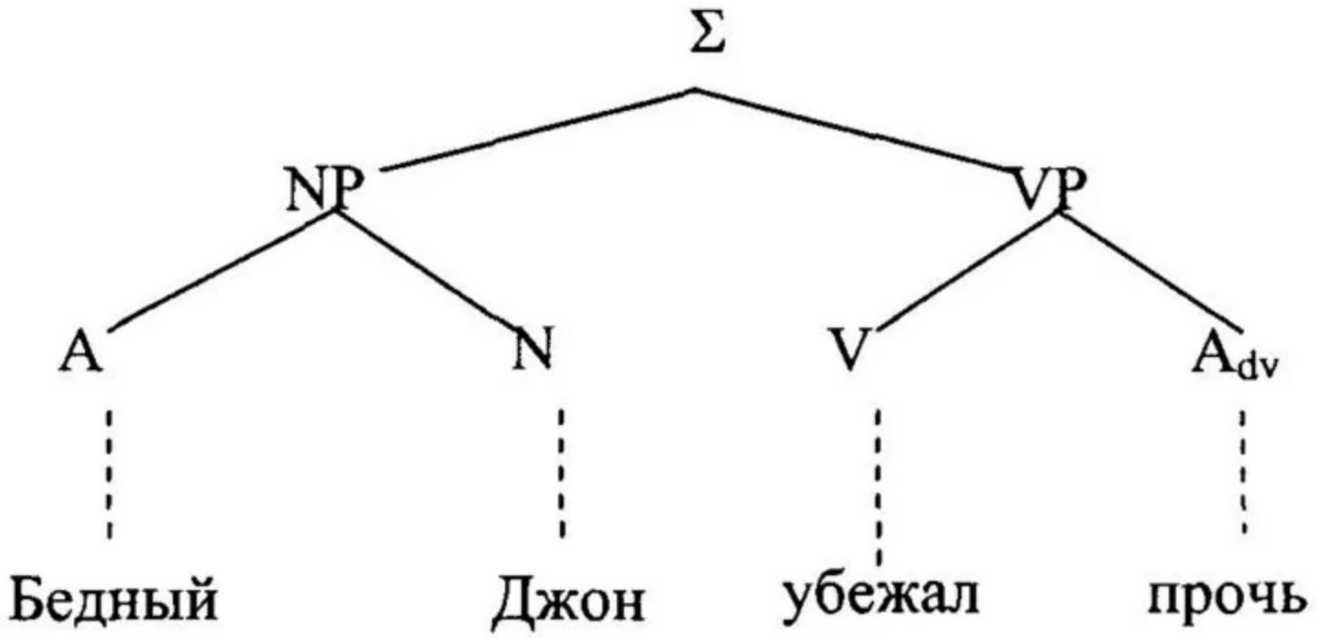
Основные виды разметок:

1. Метатекстовая
2. Морфологическая
3. Семантическая
4. Синтаксическая
5. Лемматизация

Национальный корпус русского языка

Есть несколько разметок:

- Метатекстовая
- Морфологическая
- Синтаксическая
- Семантическая



Глубоко аннотированный корпус имеет дерево зависимостей для каждого предложения

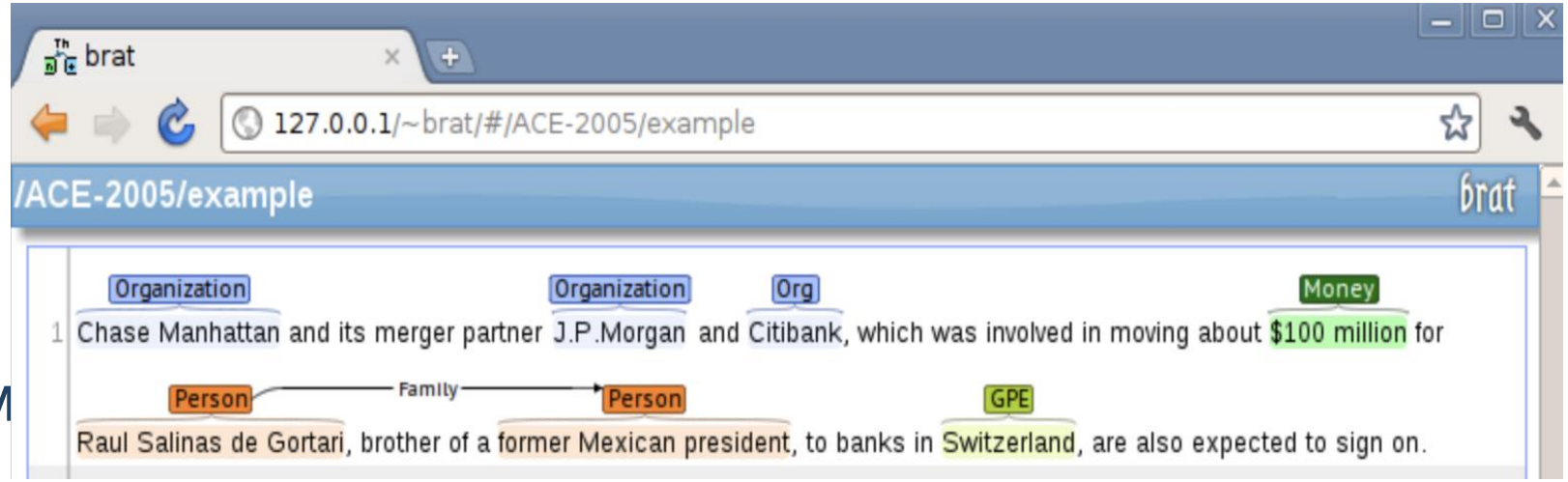
Британский национальный корпус

- Разметка содержит лемму, часть речи и синтаксическую роль
- Нет своей системы для работы с корпусом
- Можно искать через SketchEngine, English-Corpora.org, Intellitext

```
<head type="MAIN">
    <s n="1"><w c5="VBG" hw="be" pos="VERB">BEING </w>
        <w c5="VVN" hw="draw" pos="VERB">DRAWN </w>
        <w c5="PRP" hw="to" pos="PREP">TO </w>
        <w c5="AT0" hw="an" pos="ART">AN </w>
        <w c5="NN1" hw="image" pos="SUBST">IMAGE</w>
    </s>
</head>
```

Система Brat

- Система для разметки корпусов
- Можно работать параллельно над одним текстом
- Визуализирует разметку
- Предоставляет свой формат аннотации




```
T1      Catalysis 162 173       vertebrates
E1      Catalysis:T1
T2      Acetylation 531 538     species
E2      Acetylation:T2
T3      Phosphorylation 834 843 diaphragm
E3      Phosphorylation:T3
```

Система Gate

- Система для обработки естественного языка
- Основной формат документов - XML
- Файлы можно хранить в DataStores
- GATE Embedded

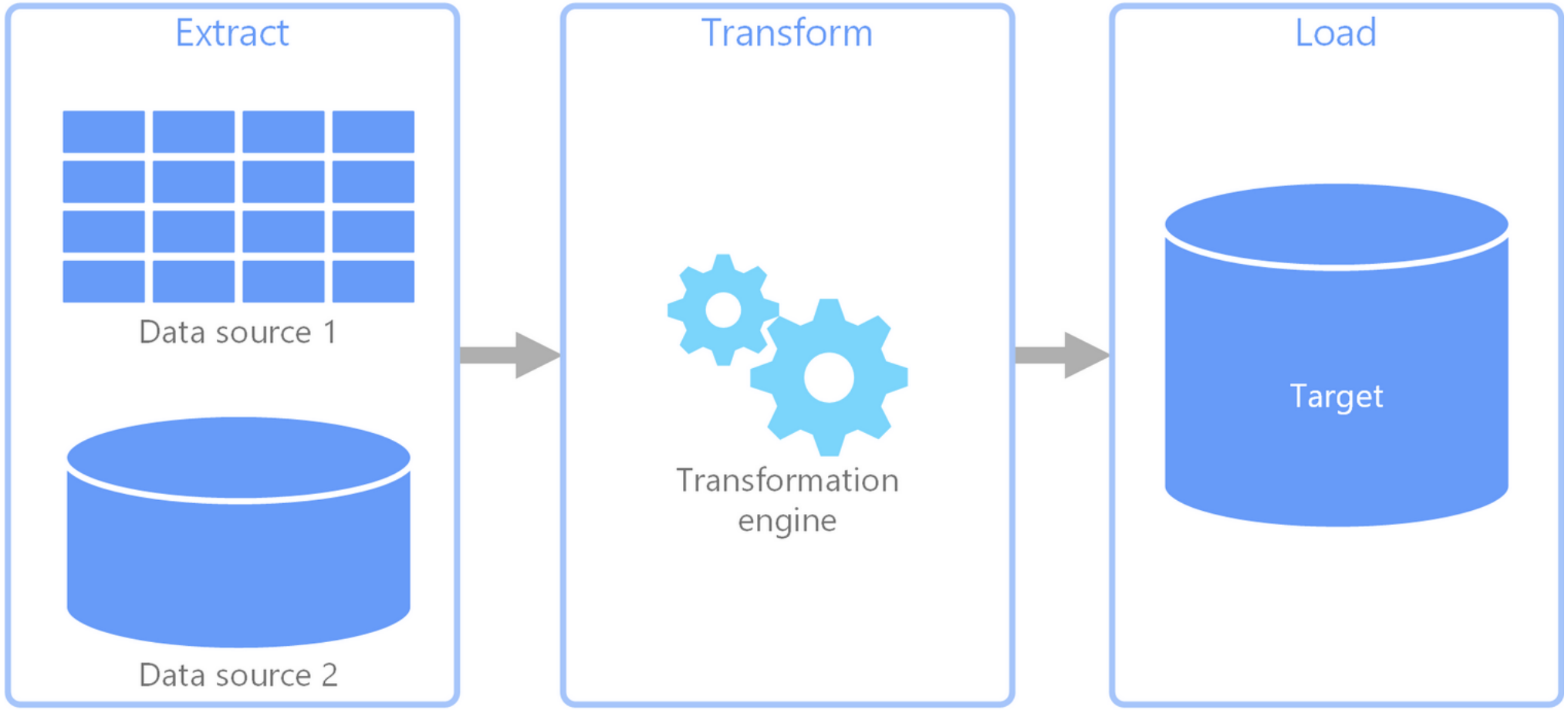
Текст – группа предложений, объединённых в одно целое темой и основной мыслью. Предложения в тексте связаны по смыслу и при помощи языковых средств (повтор, местоимения, синонимы и др.).

- firstWord
- myAnnotation

```
<TextWithNodes>
    <Node id="0"/>
    Текст
    <Node id="5"/>
    – группа предложений, объединённых в одно целое темой и основной
    <Node id="70"/>
    мыслью.
    <Node id="78"/>
    <Node id="79"/>
    Предложения
    <Node id="90"/>
    в тексте связаны по смыслу и при помощи языковых
    <Node id="140"/>
    средств
    <Node id="147"/>
    (повтор, местоимения, синонимы и др.).
</TextWithNodes>
    <!-- The default annotation set -->
<AnnotationSet> </AnnotationSet>
    <!-- Named annotation set -->
<AnnotationSet Name="firstWord"> </AnnotationSet>
    <!-- Named annotation set -->
<AnnotationSet Name="myAnnotation">
    <Annotation EndNode="5" StartNode="0" Type="firstWord" Id="1"> </Annotation>
    <Annotation EndNode="90" StartNode="79" Type="firstWord" Id="2"> </Annotation>
    <Annotation EndNode="78" StartNode="70" Type="lastWord" Id="8"> </Annotation>
    <Annotation EndNode="147" StartNode="140" Type="lastWord" Id="15"> </Annotation>
</AnnotationSet>
```

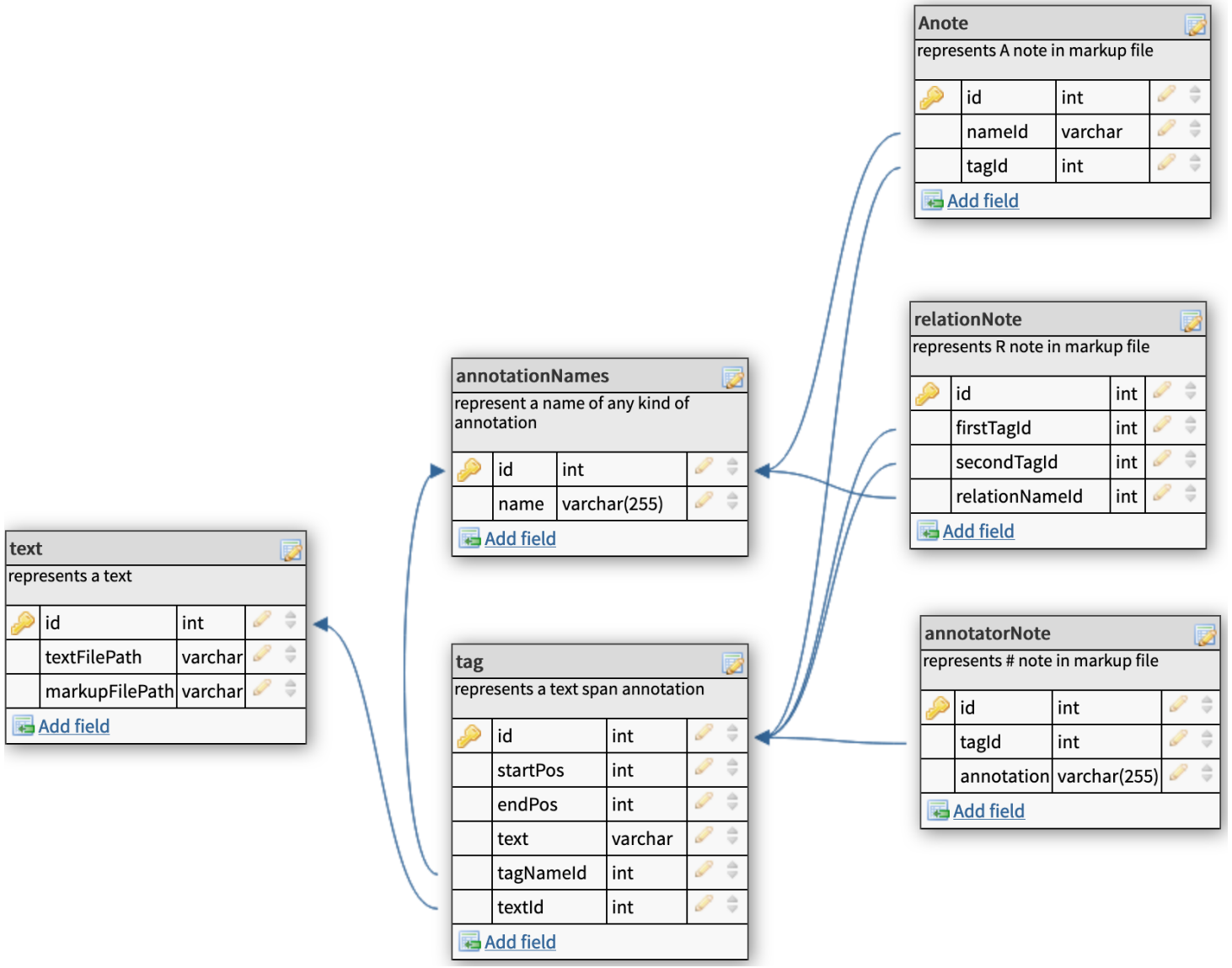
ETL системы

ETL – Extract, Transform, Load



Трансформация разметки Brat

```
T1      Catalysis 162 173       vertebrates
E1      Catalysis:T1
T2      Acetylation 531 538     species
E2      Acetylation:T2
T3      Phosphorylation 834 843 diaphragm
E3      Phosphorylation:T3
```



Декларативный механизм трансформации XSLT

```xml
<head type="MAIN">
    <s n="1"><w c5="VBG" hw="be" pos="VERB">BEING </w>
        <w c5="VVN" hw="draw" pos="VERB">DRAWN </w>
        <w c5="PRP" hw="to" pos="PREP">TO </w>
        <w c5="AT0" hw="an" pos="ART">AN </w>
        <w c5="NN1" hw="image" pos="SUBST">IMAGE</w>
    </s>
</head>
```

c5	hw	pos	word
VVN	draw	VERB	DRAWN
PRP	to	PREP	TO
AT0	an	ART	AN
NN1	image	SUBST	IMAGE

```xml
<xsl:stylesheet version="1.0" xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
    <xsl:template match="/wtext/div/head/s">
        <xsl:text> c5 | hw | pos | word </xsl:text>
        <xsl:value-of select="'&#13;'" />
        <xsl:apply-templates select="День/Занятие"/>
    </xsl:template>
    <xsl:template match="/wtext/div/head/s/w">
        <xsl:value-of select="concat(@c5, ' | ', @hw,' | ', @pos, ' | ', @word, '&#13;')"/>
    </xsl:template>
</xsl:stylesheet>
```