

# XML, XML-TEI + Python :)

Материалы Д. Скоринкина

# Что будет сегодня

- **XML** вообще и в контексте DH; TEI/XML
- Посмотрим **XML** из Питона
- Идея **xPath** для **XML**

# XML = eXtensible Markup Language

- XML — скорее **метаязык**, чем язык
- XML — это обобщенный **синтаксис** языков разметки  
**(без семантики)**
  - Конкретные языки вроде TEI — «расширения» или диалекты XML
- XML придуман, чтобы хранить и передавать данные
- XML не привязан к конкретному софту, железу или разработчику\*

\*разработку и развитие XML ведет консорциум W3C

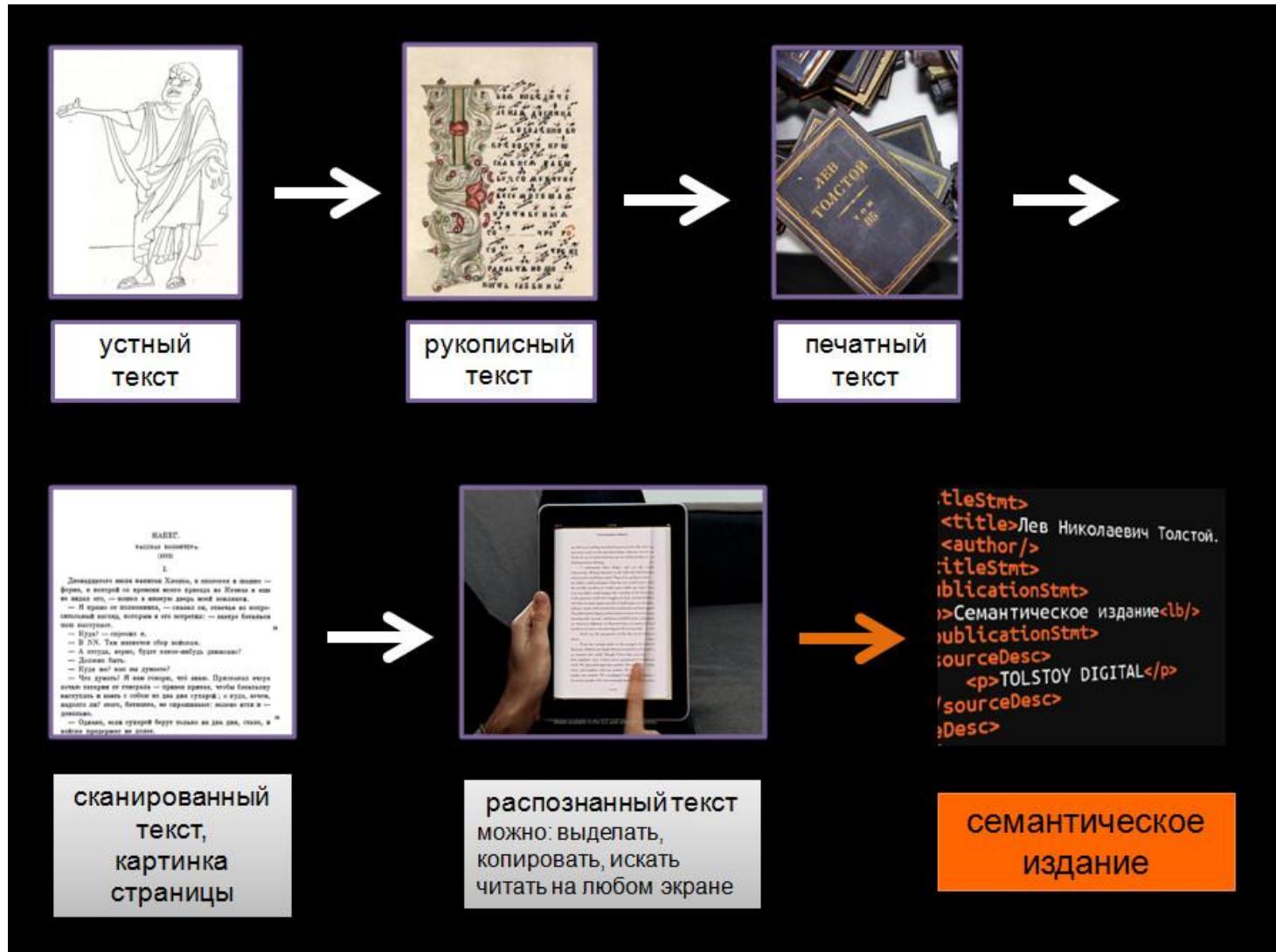
# Почему полезно посмотреть на XML?

- Это очень **распространенная технология** передачи и обмена данными\*
  - например, форматы .docx, .xlsx, .epub — это XML
- Человекочитаемость + машиночитаемость
- Оптимален для **привязки** информации к **<span>**конкретным отрезкам**</span>** текста
- На нем основан TEI , а если у вас нет TEI, в DH вас ~~уважать не будут~~

# XML в DH-контексте

В DH популярна концепция  
«академических цифровых изданий»

# Эволюция форм представления текста



```
<titleStmt>
<title>Лев Николаевич Толстой.</title>
<author/>
<titleStmt>
<publicationStmt>
<sourceDesc>
<p>TOLSTOY DIGITAL</p>
<sourceDesc>
<desc>
```

Цифровой текст — это еще не  
машиночитаемые данные, а всего  
лишь цепочка символов

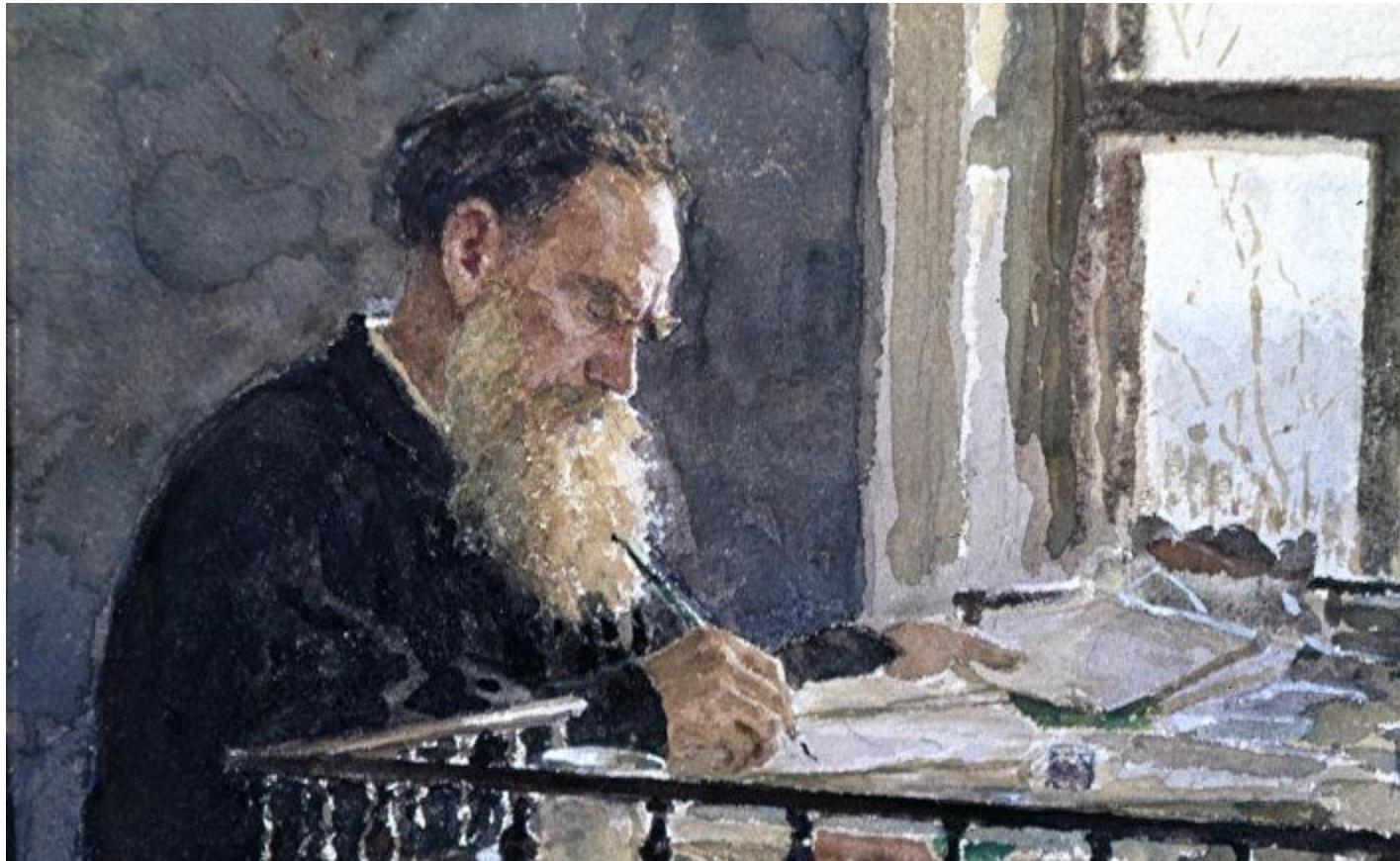
# Компьютер не умеет «понимать» тексты, даже когда они оцифрованы

Выявление в тексте сущностей, разметка структуры документа и даже простое разделение на слова — сложные интеллектуальные задачи, которые со 100% точностью не решает ни одна компьютерная система.

Андрей Болконский  
Князь Андрей, не оглядываясь, сморщил . Андрей Болконский  
гримасу, выражавшую досаду на Пьер Безухов огает его  
за руку, но, увидев улыбающееся лицо Пьера,  
Прямая речь  
Кто: Андрей Болконский  
Кому: Пьер Безухов  
— Вот как!.. И ты в большом свете! — сказал он Пьеру.  
Пьер Безухов  
— Я знал, что вы будете — отвечал Пьер. — Я приеду к  
вам ужинать, — пр не мешать  
виконту, который каз. — Можно?  
Прямая речь  
Кто: Пьер Безухов  
Кому: Андрей Болконский

Такое компьютеру пока сложно

И тем более компьютер не умеет брать текст и находить метаинформацию о нем (дата и обстоятельства написания, имя, пол и возраст автора, жанр...)



И вот однажды гуманитарии  
собрались и подумали



конференция по текстовой разметке в Вассар-колледже, ноябрь 1987

# ...подумали, что им нужна такая разметка

- которая позволит привязывать машиночитаемую информацию к тексту
- которую компьютеру не придется «понимать» — т.е. она будет однозначна
- которая будет хранить метаданные о документе вместе с самим документом
  - происхождение текста: библиография, провенанс рукописи и т.п.
  - физические свойства оригинала
  - нашу ответственность за оцифровку и принятые нами решения
- которая будет понятна и человеку

Например, такая:

Что ж мой <**persName**>Онегин</**persName**>? Полусонный

В постелю с бала едет он:

А <**placeName**>Петербург</**placeName**> неугомонный

Уж барабаном пробужден.

Или чуть посложнее:

**<author>**А. С. Пушкин**</author>**

Что ж мой **<persName**

**who=“EOnegin”>**Онегин**</persName>**? Полусонный

В постелю с бала едет **<reference**

**who=“EOnegin”>**он**</reference>**:

А **<placeName lat=“59.57” long=“30.19”**

**>**Петербург**</placeName>** неугомонный

Уж барабаном пробужден.

# Или еще сложнее:

```
<?xml-model href="https://dracor.org/schema.rng" type="application/xml" schematypens="http://relaxng.org/ns/structure/1.0"?>
▼<TEI xmlns="http://www.tei-c.org/ns/1.0" xml:lang="rus">
  ▼<teiHeader>
    ▼<fileDesc>
      ▼<titleStmt>
        <title type="main">Смерть Иоанна Грозного</title>
        <title type="main" xml:lang="eng">The Death of Ivan the Terrible</title>
        <title type="sub">Трагедия в пяти действиях</title>
        <title type="sub" xml:lang="eng">A Tragedy in Five Acts</title>
      ▼<author>
        ▼<persName>
          <forename>Алексей</forename>
          <forename type="patronym">Константинович</forename>
          <surname>Толстой</surname>
        </persName>
        ▼<persName xml:lang="eng">
          <forename>Aleksey</forename>
          <surname>Tolstoy</surname>
        </persName>
        <idno type="wikidata">Q212575</idno>
      </author>
    </titleStmt>
    ▼<publicationStmt>
      <publisher xml:id="dracor">DraCor</publisher>
      <idno type="dracor" xml:base="https://dracor.org/id/">rus000076</idno>
    ▼<availability>
      ▼<licence>
        <ab>CC0 1.0</ab>
        <ref target="https://creativecommons.org/publicdomain/zero/1.0/">Licence</ref>
      </licence>
    </availability>
    <idno type="wikidata" xml:base="http://www.wikidata.org/entity/">Q4424383</idno>
  </publicationStmt>
  ▼<sourceDesc>
    ▼<bibl type="digitalSource">
      <name>Wikisource</name>
      <idno type="URL">https://ru.wikisource.org/wiki/Смерть_Иоанна_Грозного_(А._К._Толстой)</idno>
    ▼<availability>
      ▼<licence>
        <ab>CC BY-SA 3.0</ab>
        <ref target="https://creativecommons.org/licenses/by-sa/3.0/deed.ru">Licence</ref>
      </licence>
    </availability>
  </sourceDesc>

```

# Главные правила XML

XML — это всегда дерево:

<text>

<line>

<word>Нет</word> <word>войне</word>

</line>

</text>

Теги XML *обязательно* должны закрываться

Не может быть тегов «вперехлест» (каждый тег должен быть вложен в какой-то другой; корневой тег один)

Так нельзя:

<text>

<line>

<word>Нет</word> <word>войне</word>

</text>

</line>



непонятно, кто в кого входит (line в text? text в line?)

# И так нельзя, если это целый документ

<text>

Digital

</text>

<text>

Humanities



</text>

корневой тег должен быть один!

# Если тег пустой...

<text>

<line>

<word>Нет</word> <word>войне</word>

</line>

<line></line>

</text>

Если тег пустой, его можно сделать  
самозакрывающимся

<text>

<line>

<word>Нет</word> <word>войне</word>

</line>

<line/>

</text>



А так — нельзя (ср. HTML, где так вполне можно)

<text>

<line>

<word>Нет</word> <word>войне</word>

</line>

<line>

</text>



# У элементов XML бывают атрибуты

```
<text id="001">  
    <line number="1">Нет</line>  
    <line number="2">войне</line>  
</text>
```

`@id, @number` — атрибуты этого XML

1,2 — значения (values) атрибута `@number`

Значения всегда в кавычках!

## «Драконовский контроль ошибок» (draconian error handling)

- XML-документ, не соблюдающий правила XML-я (не ‘well-formed’), не считается XML-документом
- С большой вероятностью программы, предназначенные для работы с XML, откажутся обрабатывать такой документ и выдадут ошибку

Что добавляет к этому TEI?

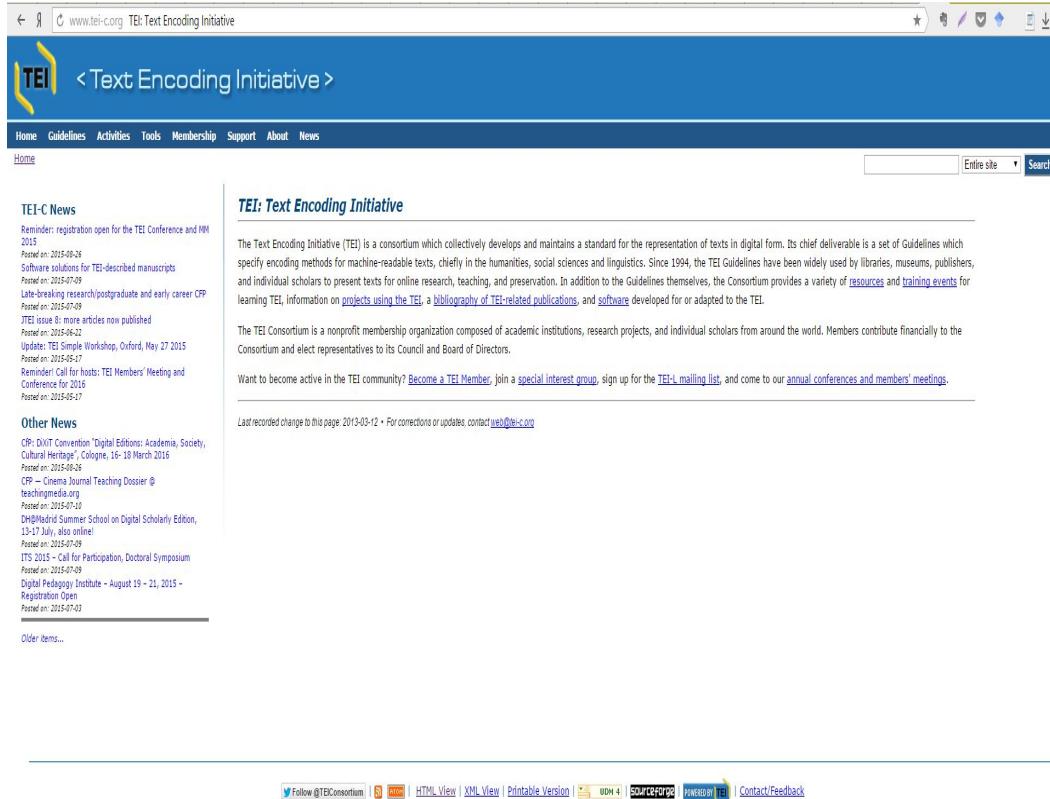
# TEI (Text Encoding Initiative)



- *Расширение XML с конкретными тегами*, о значении которых договорилось большое сообщество людей
- «Мировой стандарт разметки для гуманитариев»: предписывает, как единообразно кодировать поэзию, прозу, драму, цифровые коллекции рукописей, писем, словари\* и лингвистические корпуса\*...
- Инициативное сообщество, разрабатывающее этот стандарт

\*тут TEI пока совсем не закрепился как стандарт

# Text Encoding Initiative Consortium



The screenshot shows the homepage of the Text Encoding Initiative (TEI) website. At the top, there's a navigation bar with links for Home, Guidelines, Activities, Tools, Membership, Support, About, and News. Below the navigation is a search bar with fields for 'Entire site' and 'Search'. The main content area features several news items under 'TEI-C News' and 'Other News'. One news item is highlighted: 'Reminder: registration open for the TEI Conference and MM 2015' (Posted on: 2015-08-24). Other news items include 'Software solutions for TEI-described manuscripts' (Posted on: 2015-07-01), 'Late-breaking research/postgraduate and early career CFP' (Posted on: 2015-07-09), 'JTEI issue 8: more articles now published' (Posted on: 2015-06-32), 'Update: TEI Simple Workshop, Oxford, May 27 2015' (Posted on: 2015-05-17), 'Reminder! Call for hosts: TEI Members' Meeting and Conference for 2016' (Posted on: 2015-05-17), and 'CIP: DIXIT Convention "Digital Editions: Academia, Society, Cultural Heritage", Cologne, 16- 18 March 2016' (Posted on: 2015-08-28). There's also a mention of 'CIP – Cinema Journal Teaching Dossier' (Posted on: 2015-07-08), 'DH@Hybrid Summer School on Digital Scholarly Edition, 13-17 July, also online' (Posted on: 2015-07-09), 'ITS 2015 - Call for Participation, Doctoral Symposium' (Posted on: 2015-07-09), 'Digital Pedagogy Institute - August 19 - 21, 2015 - Registrations Open' (Posted on: 2015-07-02), and a link to 'Older items...'. At the bottom of the page, there are social media links for Twitter, Facebook, and Google+, along with links for 'HTML View', 'XML View', 'Printable Version', 'UDH 4', 'SourceForge', 'PINSER', and 'Contact/Feedback'.

- 71 институт, сотни индивидуальных участников
- воркшопы, школы, конференции
- Journal of the Text Encoding Initiative
- 1636 страниц guidelines (в pdf-версии)
- примеры, образцы, мультиязычность, каталоги ресурсов

www.tei-c.org

# Базовые возможности TEI

- Версии и разнотечения внутри одного текста
- Структура документа
- Хранение метаданных без отрыва от документа
- Encoding не только Text

## Версии одного текста

Меня преследуют две-три случайных фразы,

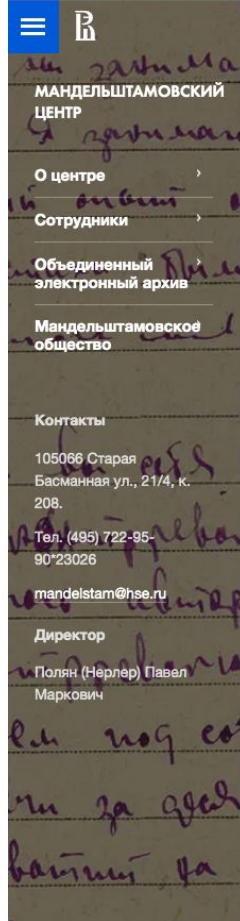
Весь день твержу: печаль моя жирна

О Боже, как жирны и синеглазы

Стрекозы смерти, как лазурь черна.

О. Мандельштам

# Версии одного текста



Национальный исследовательский университет «Высшая школа экономики» → Учебные подразделения → Факультет гуманитарных наук → Департамент общей и прикладной филологии → Мандельштамовский центр → Архив → 10 января 1934 года (Памяти Б.Н.Бугаева). "Воспоминание". Машинопись. Л.1-4. [1930-е]



## 10 января 1934 года (Памяти Б.Н.Бугаева). "Воспоминание". Машинопись. Л.1-4. [1930-е]

Объединенный  
электронный  
архив Осипа  
Мандельштама

Описание Ф. 613 (ГИХЛ). Оп. 1. Д. 4686. Л. 1-4.

Листы 4 л.

Автор Мандельштам О.Э.

Дата 1930-е



# Версии одного текста

ОСИД МАНДЕЛЬСТАМ

10 января 1934 года

Памяти Б. Н. Бугаева /Андрея  
Белого/

Мои проследует две, три случайных фразы,  
жирна  
Зесь донъ творжу: печаль моя жирна  
О, боже, как жирна и синеглава  
Стрекозы смерти, как лазурь черна  
Где первородство? где счастливая повалка?  
планки  
Где пижиний котребок на самом дне очей  
Где величество? где горькая украдка?  
прямизна  
Где ясный стан? где хххххх речей?

## Версии одного текста

Меня преследуют две-три случайных фразы,  
жирна

Весь день твержу: печаль моя ~~жарка~~

О Боже, как жирны и синеглазы

Стрекозы смерти, как лазурь черна.

О. Мандельштам

## Версии одного текста

Меня преследуют две-три случайных фразы,

Весь день твержу: печаль моя

жирна

жарка

О Боже, как жирны и синеглазы

Стрекозы смерти, как лазурь черна.

О. Мандельштам

## Версии одного текста в TEI

Меня преследуют две-три случайных фразы,  
Весь день твержу: печаль моя <choice>  
                  <del>жарка</del>  
                  <add>жирна</add>  
                  </choice>

О Боже, как жирны и синеглазы  
Стрекозы смерти, как лазурь черна.

## Версии одного текста в TEI

**<choice>** — создаем разветвку из нескольких вариантов

**<del>** — что было удалено

**<add>** — что было добавлено

## Версии одного текста

зуется ложей.) А рябая родственница сказала: «верхомъ бы въ мискѣ еще тепленъкій довезли. Вотъ Княгиня И. В. всегда посыпаетъ верхомъ<sup>2</sup> да еще къ Сухаревой башнѣ». Всѣ обратили вниманіе на рябую родственницу. — «Полноте, М. И.», сказалъ, добродушно улыбаясь, папа.

<sup>2</sup> Написано: вихремъ.

## Версии одного текста в TEI

Вотъ Княгиня И. В. всегда посылаетъ

<choice>

<sic>вихремъ</sic>

<corr resp="#editor1">верхомъ</corr>

</choice>

да еще къ Сухаревой башнъ

## Версии одного текста в TEI

**<choice>** — создаем разветвку из нескольких вариантов

**<sic>** — так у автора

**<corr>** — исправленный вариант

# Структура документа

Две первые строфы того же стихотворения О. Мандельштама:

```
<lg type="quatrains">
    <l>Меня преследуют две-три случайных фразы,</l>
    <l>Весь день твержу: печаль моя жирна.</l>
    <l>О, Боже, как жирны и синеглазы</l>
    <l>Стрекозы смерти, как лазурь черна!</l>
</lg>
<lg type="quatrains">
    <l>Где первородство? Где счастливая повадка?</l>
    <l>Где плавкий ястребок на самом дне очей?</l>
    <l>Где вежество? Где горькая украдка?</l>
    <l>Где ясный стан? Где прямизна речей,</l>
</lg>
```

## Структура документа: поэзия

**<|>** — любая стихотворная строка

**<lg>** — любое объединение стихотворных строк (например, строфа)

**@type** — атрибут «тип элемента»

# Структура документа: проза (тома — главы — части)

Начало «Войны и мира» в TEI:

```
<div n="1" type="volume" xml:id="Volume1">
  <div n="1" type="part" xml:id="part1Volume1">
    <div n="1" type="chapter" xml:id="chapter1part1Volume1">
      <p>
        <s>Eh bien, mon prince. Gênes et Lueques ne sont plus que des ap
          de la famille Buonaparte. Non, je vous préviens que si vous ne
          avons la guerre, si vous vous permettez encore de pallier tout
          toutes les atrocités de cet Antichrist (ma parole, j'y crois)
          plus, vous n'êtes plus mon ami, vous n'êtes plus мой верный ра
          dites.</s><s>Ну, здравствуйте, здравствуйте. Je vois que je vo
          и рассказывайте.</s></p>
      <p><s>Так говорила в июле 1805 года известная Анна Павловна Шерер,
        приближенная императрицы Марии Феодоровны, встречая важного и
        Василия, первого приехавшего на ее вечер.</s>
```

## Структура документа: проза

**<div>** — любая единица членения прозаического текста

**@type** — атрибут «тип элемента»

**<p>** — абзац

**<s>** — строка (в прозе)

# Структура документа: драма (сцены — явления)

```
<div type="act">
    <head>Действие первое</head>
    <stage>Комната в доме городничего.</stage>
    <div type="scene">
        <head>Явление I</head>
        <stage>Городничий, попечитель богоугодных заведений,
            смотритель училищ, судья, частный
            пристав, лекарь, два квартальных.</stage>
        <sp who="#gorodnichij">
            <speaker>Городничий.</speaker>
            <p>Я пригласил вас, господа, с тем чтобы сообщить вам
                пренеприятное известие: к нам едет ревизор.</p>
        </sp>
        <sp who="#ammos_fedorovich_ljapkin_tjapkin">
            <speaker>Аммос Федорович.</speaker>
            <p>Как ревизор?</p>
        </sp>
```

## Структура документа: драма

**<sp>** — реплика в драме

**@who** — атрибут «ID персонажа»

**<speaker>** — упоминание персонажа,  
произносящего реплику

**<stage>** — сценическая ремарка

# Метаинформация без отрыва от документа

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title type="main" xml:lang="ru">Ревизор</title>
      <title type="main" xml:lang="en">The Government Inspector</title>
      <title type="sub" xml:lang="ru">Комедия в пяти действиях</title>
      <title type="sub" xml:lang="en">A Comedy in Five Acts</title>
      <author key="wikidata:Q43718">Гоголь, Николай Васильевич</author>
    </titleStmt>
    <publicationStmt> [10 lines]
    <sourceDesc>
      <bibl type="digitalSource">
        <name>Интернет-библиотека Алексея Комарова</name>
        <idno type="URL">http://ilibrary.ru/text/473/index.html</idno>
        <availability status="free">
          <p>In the public domain.</p>
        </availability>
      <bibl type="originalSource">
        <title>Н. В. Гоголь. Собрание сочинений в 9 т. Т. 4. М.: Русская книга, 1994.</title>
        <date type="print" when="1836">"Дата первой публикации: 1836" (Wikipedia)</date>
        <date type="premiere" when="1836">"Первые представления шли в первой редакции 1836 года." (Wikipedia)</date>
```

# Метаинформация в TEI

**<teiHeader>** — элемент со всеми метаданными

**@who** — атрибут «ID персонажа»

**<speaker>** — упоминание персонажа, произносящего реплику

**<stage>** — сценическая ремарка

# Метаинформация без отрыва от документа

```
<fileDesc>
  <titleStmt>
    <title> Война и мир. </title>
    <author> Толстой Л.Н. </author>
    <respStmt>
      <resp> подготовка электронного издания </resp>
      <name> Даниил Скоринкин </name>
    </respStmt>
  </titleStmt>
  <publicationStmt>
    <publisher>
      <orgName> Центр цифровых гуманитарных исследований НИУ ВШЭ </orgName>
    </publisher>
    <availability>
      <p> Тексты: © Электронная публикация – РВБ, 2002–2019.  

        Метатекстовая разметка доступна для свободного использования и  

        распространения по лицензии Creative Commons Attribution Share-Alike (cc by-sa) </p>
    </availability>
  </publicationStmt>
  <sourceDesc>
    <bibl type="digitalSource">
      <name>Русская виртуальная библиотека (РВБ)</name>
      <idno type="URL">https://rvb.ru/tolstoy/toc.htm</idno>
      <biblStruct type="originalSource">
        <analytic>
          <author> Толстой Л.Н. </author>
```

# Метаинформация без отрыва от документа

```
<sourceDesc>
  <bibl type="digitalSource">
    <name>Русская виртуальная библиотека (РВБ)</name>
    <idno type="URL">https://rvb.ru/tolstoy/toc.htm</idno>
    <biblStruct type="originalSource">
      <analytic>
        <author> Толстой Л.Н. </author>
        <title level="a"> Война и мир. </title>
      </analytic>
      <monogr>
        <title level="m"> Собрание сочинений в 22 томах. Тома 4 - 7 </title>
        <imprint>
          <pubPlace> Москва </pubPlace>
          <publisher> "Художественная литература" </publisher>
          <date when="1979"/>
        </imprint>
      </monogr>
      <series>
        <title level="s"> Л.Н. Толстой. Собрание сочинений в 22 томах</title>
        <biblScope unit="vol"> 4 </biblScope>
        <biblScope unit="vol"> 5 </biblScope>
        <biblScope unit="vol"> 6 </biblScope>
        <biblScope unit="vol"> 7 </biblScope>
      </series>
    </biblStruct>
  </bibl>
</sourceDesc>
```

# Не только тексты



# Метаданные открытки в TEI

```
<teiHeader>
  <fileDesc>
    <titleStmt>
      <title>San Antonio River : digital edition of card 19800726_001 from the Virgolos
          collection</title>
    </titleStmt>
    <publicationStmt>
      <p>Demonstration at DH OXSS 2013</p>
    </publicationStmt>
    <sourceDesc>
      <bibl>
        <title level="m">San Antonio River (postcard)</title>
        <publisher>School Mart</publisher>
        <pubPlace>1812 South Press, San Antonio, Texas 70210</pubPlace>
        <idno>SA-146-C</idno>
        <note resp="#ed">The San Antonio river, often called the Venice of Texas, winds its way
          through the business section of San Antonio. It is very picturesque with its many
          bridges and beautifully landscaped banks.</note>
      </bibl>
    </sourceDesc>
  </fileDesc>
</teiHeader>
```

# Описание изображения и текста открытки в TEI

```
<div type="recto">
  <figure>
    <graphic url="../../Graphics/Cartes/19800726_001r.jpg"/>
    <figDesc>View of a stream with a stone bridge and little mexican-style houses. In the
      foreground a man and a woman are riding in an open boat.</figDesc>
    <head>San Antonio River</head>
  </figure>
</div>
<div facs="19800726_001v.jpg" type="verso">
  <div type="message" xml:lang="fr">
    <p>
      <date when="1980-07-26">26 juill 80</date>
    </p>
    <p>Chère Madame, après New-York et Washington dont le gigantisme m'a beaucoup séduite,
      nous avons commencé notre conquête de l'Ouest par New Orleans, ville folle en fête
      perpétuelle. Il fait une chaleur torride au Texas mais le coca-cola permet de résister -
      l'Amérique m'enchanté ! Bientôt, le grand Canyon, le Colorado et San Francisco... </p>
    <p> En espérant que vous passez de bonnes vacances, affectueusement. </p>
    <signed>Sylvie </signed>
    <signed>François </signed>
  </div>
  <div type="destination"> [20 lines]
</div>
```

# «Почтовые» метаданные

```
<div type="destination">
  <ab>
    <stamp type="postmark">
      <placeName>El Paso</placeName> - TX 799 -<date notBefore="1980-07-26">
        <unclear>PM JUL</unclear>
      </date>
    </stamp>
    <stamp type="postage">Profil masculin, avec un avion et un radar au second plan:
      <mentioned>US Airmail 21 c.</mentioned>
    </stamp>
  </ab>
  <ab>
    <address>
      <addrLine>Madame <name>Lefrère</name>
      </addrLine>
      <addrLine>4, allée George Rouault</addrLine>
      <addrLine>75020 Paris</addrLine>
      <addrLine>France</addrLine>
    </address>
  </ab>
</div>
```

# ! XML (и TEI/XML) — не язык программирования. Он сам по себе ничего не делает

XML просто кодирует информацию  
в формальном виде:

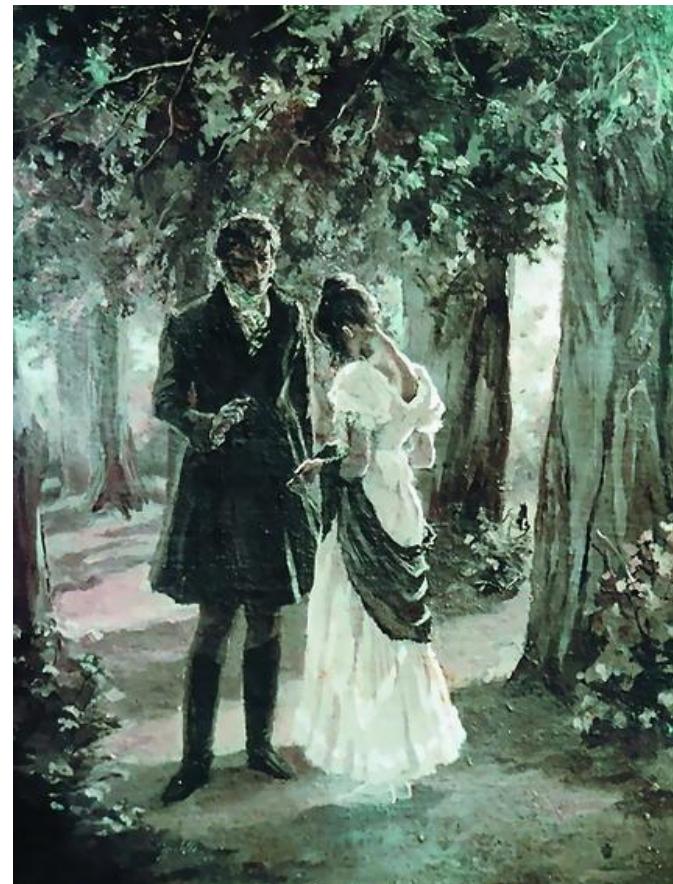
<message>

<recipient>Онегин</recipient>

<sender>Татьяна</sender>

<body>Я к вам пишу – чего же боле?</body>

</message>



[Иллюстрации Лидии Тимошенко к Евгению Онегину](#)

# Для работы с XML есть много инструментов

- Есть целая экосистема инструментов вокруг XML.  
Проще всего узнать о ней тут:  
[w3schools.com/xml](https://www.w3schools.com/xml)
- Например, можно задавать запросы к данным в XML  
(**xPath/xQuery**)
- А можно написать программу, которая превратит XML  
в нужный нам HTML/EPUB/PDF
  - для этого есть **XSLT**
  - либо с помощью **библиотек для работы с XML** в  
языках программирования, например, в **Python**

# Именованные сущности и даты

<p>

~~Eh bien, mon prince. <placeName>Gênes</placeName> et <placeName>Lueques</placeName> ne sont plus que des apanages, des поместья, de la famille Buonaparte.  
Non, je vous préviens que si vous ne me dites pas que nous avons la guerre, si vous vous permettez encore de pallier toutes les infamies, toutes les atrocités de cet Antichrist (ma parole, j'y crois) - je ne vous connais plus, vous n'êtes plus mon ami, vous n'êtes plus мой верный раб, comme vous dites.</s><s>Hy, здравствуйте, здравствуйте. Je vois que je vous fais peur, садитесь и рассказывайте.</s></p>~~

<p>

~~Так говорила в <date when="1805-06">июле 1805 года</date> известная <persName ref="Anna\_Pavlovna\_Scherer">Анна Павловна Шерер</persName>, фрейлина и приближенная <persName ref="empress\_Mariya"> <roleName type="title">императрицы</roleName> Марии Феодоровны</persName>, встречая важного и чиновного <persName ref="Vasili\_Kuragin"> <roleName type="title">князя</roleName> Василия</persName>, первого приехавшего на <rs ref="Anna\_Pavlovna\_Scherer">ее</rs> вечер.</s>~~

</p>

# Именованные сущности и даты в TEI

**<placeName>** — топоним (может иметь в качестве атрибутов широту и долготу)

**<persName>** — имя человека

**<rs>** — referring string, любая часть текста, которая является ссылкой на именованной сущности (исп. вместе с ID сущности)

**<date>** — время (записывается в стандартизированном формате YYYY-MM-DD)

# Поэзия (Verse) — метрическая структура

```
<lg type="quatrains" rhyme="AbAb" met="-+|-+|-+|-+|/-+|-+|-+| -+>
<l real="-+|-+|-+|-+| -">Мой дядя самых честных правил, </l>
<l real="-+|-+|-+|-+| -">Когда не в шутку занемог, </l>
<l real="--|-+|-+|-+| -">Он уважать себя заставил</l>
<l real="-+|-+|-+|-+| -">И лучше выдумать не мог.</l>
</lg>
```

Более точно:

```
<lg type="quatrains" rhyme="AbAb" met="-+|-+|-+|-+|/-+|-+|-+| -+>
<l>Мой дядя самых честных правил, </l>
<l>Когда не в шутку <seg type="foot" n="3" real="--">зане</seg>мог, </l>
<l><seg type="foot" n="1" real="--">Он у</seg>важать себя заставил</l>
<l>И лучше вы<seg type="foot" n="3" real="--">думатель</seg> не мог.</l>
</lg>
```

# Поэзия (Verse) — зона рифмовки

```
<lg type="quatrains">
<l>Мой дядя самых честных <rhyme>правил</rhyme>, </l>
<l>Когда не в шутку зане<rhyme>мог</rhyme>, </l>
<l>Он уважать себя за<rhyme>ставил</rhyme></l>
<l>И лучше выдумать не <rhyme>мог</rhyme>. </l>
</lg>
```

# Критический аппарат (critical apparatus)

и маху въ Варази изъ заморья на Чюди и на Сло  
вѣнех на Мери и на всѣхъ Кривичѣхъ а Козари  
имаху на Поланѣхъ и на Сѣверѣхъ на Вятичѣхъ им[а]  
ху по бѣлѣи вѣверицѣ ѿ дыма

[И]маху Въ дань Варази изъ заморья на Чюди и на Сло  
вѣнех на Мери и на всѣхъ Кривичѣхъ а Козари  
имаху на Поланѣхъ и на Сѣверѣхъ на Вятичѣхъ им[а]  
ху по бѣлѣи вѣверицѣ ѿ дыма

Лаврентьевская летопись 1377, 7 (859)

# Критический аппарат (critical apparatus)

и магчадль в азъ ппозаморыя наупдии на  
вѣнчиме и ппистлаки на путѧ козары  
магчил пла не тѣ пписте вѣт ппават путь им  
жѹп вѣлѣт вѣрицъ ѿдьма:

[И]маху В дань Ваази изъ заморья на Чюди и на Сло  
вѣнех на Мери и на всѣхъ Кривичехъ а Козари  
имаху на Поланѣхъ и на Сѣверѣхъ на Вятичехъ им[а]  
ху по **бѣлѣи** вѣверицъ ѿдьма  
**бѣлѣи**

Лаврентьевская летопись 1377, 7 (859)

# Критический аппарат (critical apparatus) в TEI

```
<app>
    <lemma wit="#w1 #w2">бълъи</lemma>
    <rdg wit="#w3">бълъ и</rdg>
</app>
```

# Критический аппарат в TEI

**<app>** — одна запись критического аппарата

**<rdg>** — одно из прочтений

**<lemma>** — доминирующее прочтение (при наличии)

**@wit** — ссылка на источник прочтения

# Рукописи (Manuscript Description)



# Метаданные рукописи в TEI

```
<msDesc>
  <msIdentifier>
    <!-- идентификатор рукописи, основные библиографические сведения-->
    <settlement>Санкт-Петербург</settlement>
    <repository>Российская национальная библиотека</repository>
    <!--место хранения-->
    <idno> F.IV.2.</idno>
    <!--шифр в инвентарном описании-->
  </msIdentifier>
  <msContents>
    <!--описание содержательной части-->
    <msItem>
      <author>монах <persName ref="#Lavrentiy">Лаврентий</persName></author>
      <!--автор/составитель текста, в атрибуте ref ссылка на описание персоны в другой части
      <textLang mainLang="orv">Древнерусский язык</textLang>
      <!--языки текста-->
    </msItem>
  </msContents>
```

# Физические свойства рукописи в TEI

```
<physDesc><!-- описание физических свойств рукописи-->
<objectDesc form="codex">
  <supportDesc material="perg">
    <!-- физические свойства носителя текста-->
    <support><!-- описание материалов и иных физических составляющих-->
      <p><material>Пергаменный</material> кодекс, присутствуют более поздние
        <material>бумажные</material> вставки</p>. </support>
    <collation> <!-- описание того, как соединены между собой страницы или
      иные части манускрипта-->
      <p>Число тетрадей - 24. Регулярное число листов в тетради - по 8 листов (имеются
        отклонения), тетрадная формула: <formula>1[8], 2[2], 3[8]-5[8], 6[6], 7[8]-20[8],
        21[12], 22[4], 23[2], 24[3] </formula>.<!-- это специальный элемент
        для тетрадной формулы--><p>
      </p>Регулярный способ
      складывания листов в тетрадях - без переворота листов, то есть на развороте
      соседствуют мясная и волосяная стороны во всех случаях, кроме центра тетради,
      где на развороте обе стороны <material>мясные</material>, и при соседстве
      тетрадей, где на развороте обе стороны <material>волосяные</material>.</p>
    </collation>
```

# Физические свойства рукописи в TEI

```
</collation>
<extent>173 листа <!--описание объема и физических размеров рукописи-->
  <dimensions scope="all" type="leaf" unit="mm"><!--описание размеров рукописи-->
    <height min="250" max="254">Высота листа – от 250 до 255 миллиметров</height>
    <width min="205" max="210"> Ширина листа – от 205 до 210 миллиметров</width>
  </dimensions>
</extent>
</supportDesc>
<bindingDesc>
  <p>Переплет по конструктивным особенностям датируется <date notBefore="1470"
    notAfter="1530">концом XV – началом XVI в.</date> В <date notBefore="1770"
    notAfter="1900">конце XVIII – XIX в.</date> блок книги был отреставрирован – многие
  листы подклеены пергаменом, заклеены некоторые отверстия, продублированы пергаменными
  подклейками многие швы, так как нити, которыми они были сделаны, в большинстве случаев
  истлели. Переплет был отреставрирован, <date notBefore="1811">когда книга уже находилась
  в <orgName>Императорской Публичной библиотеке</orgName></date>. Блок книги был
  сброшюрован, заменено крепление переплета и корешок (при этом оказались утраченными
  некоторые конструктивные особенности деревянных крышечек), на внутренние стороны крышек
  выклеены форзацы из мраморной бумаги и вставлены защитные бумажные листы. </p>
</bindingDesc>
```

# Физические свойства рукописи в TEI

```
<layoutDesc>
    <!-- описание структуры листа-->
    <layout columns="1" ruledLines="31 32">
        <!-- элемент описания структуры; в атрибутах информация о количестве колонок (@columns,
            <p>На <locus from="1" to="40"> листах 1 - 40 </locus> (тетради 1 - 6), текст написан
                уставом в одну колонку, разлиновка твердым заостренным предметом (шильцем); система -
                по два листа по волосяной стороне </p>
        </layout>
        <layout columns="2" ruledLines="31 33">
            <!-- элемент описания структуры листа-->
            <p>На <locus from="41" to="173">листах 41 - 173 </locus> текст написан русским
                полууставом в две колонки, разлиновка твердым заостренным предметом (шильцем); систем
                - Leroy 11 (без учета сторон пергамена) - по два листа: внешние два листа тетради - п
                волосяной стороне, внутренние два листа тетради - по мясной. </p>
        </layout>
    </layoutDesc>
```

# Кто приложил руку к манускрипту

```
<handDesc hands="2">
    <!--элемент для перечисления всех различных почерков или писцов;
        строго говоря, Лаврентьевская летопись содержит рукописный текст
        гораздо большего количества людей, здесь мы сознательно идем на упрощение -->
    <handNote xml:id="1" scope="minor" medium="ink, cinnabar">
        <p>Первая часть летописи (<locus from="1" to="40"> листы с 1 по 40 </locus>) написана
            неизвестным писцом, текст написан уставом</p>
    </handNote>
    <handNote xml:id="2" scope="major" scribe="#Lavrentiy" medium="ink, cinnabar">
        <p>Вторая часть летописи (<locus from="41" to="173">листы 41 - 173 </locus>) написана
            монахом <persName ref="#Lavrentiy">Лаврентием</persName>, текст написан русским
            полууставом</p>
    </handNote>
</handDesc>
```

# Навигация и запросы к данным в XML: XPath

- Язык навигации по XML-документу (применим и к большинству современных HTML-документов)

XPath 2.0 //\*

zagoskin-blagorodnyj-teatr.xml teachers.xml teachers\_bare.xml

Запрос

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3   <person>Нина Евгеньевна Сахарова</person>
4   <person>Борис Валерьевич Орехов</person>
5   <person>Даниил Андреевич Скоринкин</person>
6   <person>Франк Куртович Фишер</person>
7   <person>Георгий Алексеевич Мороз</person>
8   <person>Ольга Николаевна Ляшевская</person>
9 </teachers>
```

Выдача

Text

Grid

Description - 6 items	XPath location
Нина Евгеньевна Сахарова	/teachers[1]/person[1]
Борис Валерьевич Орехов	/teachers[1]/person[2]
Даниил Андреевич Скоринкин	/teachers[1]/person[3]
Франк Куртович Фишер	/teachers[1]/person[4]
Георгий Алексеевич Мороз	/teachers[1]/person[5]
Ольга Николаевна Ляшевская	/teachers[1]/person[6]

XPath 2.0 `/teachers/person[1]`

zagoskin-blagorodnyj-teatr.xml ● teachers.xml ● teachers\_bare.xml ●

teachers person

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3   <person>Нина Евгеньевна Сахарова</person>
4   <person>Борис Валерьевич Орехов</person>
5   <person>Даниил Андреевич Скоринкин</person>
6   <person>Франк Куртович Фишер</person>
7   <person>Георгий Алексеевич Мороз</person>
8   <person>Ольга Николаевна Ляшевская</person>
9 </teachers>
```

Text

Description - 1 item	XPath location
Нина Евгеньевна Сахарова	/teachers[1]/person[1]

XPath 2.0  //\*

zagoskin-blagorodnyj-teatr.xml  teachers.xml  teachers\_bare.xml

teachers person

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3   <person>Нина Евгеньевна Сахарова</person>
4   <person>Борис Валерьевич Орехов</person>
5   <person>Даниил Андреевич Скоринкин</person>
6   <person>Франк Куртович Фишер</person>
7   <person>Георгий Алексеевич Мороз</person>
8   <person>Ольга Николаевна Ляшевская</person>
9 </teachers>
```

Text

Description - 1 item	XPath location
Даниил Андреевич Скоринкин	/teachers[1]/person[3]

XPath 2.0 //\* ⚙ ✓ ▶ ✖ ✖ ✖ ✖ ✖ ✖ ✖

zagoskin-blagorodnyj-teatr.xml • teachers\_full.xml • teachers\_with\_attrs.xml

```

teachers person firstname
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3 <person gender="F">
4   <firstname>Нина</firstname>
5   <patronym>Евгеньевна</patronym>
6   <surname>Сахарова</surname>
7 </person>
8 <person gender="M">
9   <firstname>Борис</firstname>
10  <patronym>Валерьевич</patronym>
11  <surname>Орехов</surname>
12 </person>
13 <person gender="M" birthdate="1989-12-10">
14   <firstname>Даниил</firstname>
15   <patronym>Андреевич</patronym>

```

Text Grid Auth

Description - 6 items	XPath location
Нина	/teachers[1]/person[1]/firstname[1]
Борис	/teachers[1]/person[2]/firstname[1]
Даниил	/teachers[1]/person[3]/firstname[1]
Франк	/teachers[1]/person[4]/firstname[1]
Георгий	/teachers[1]/person[5]/firstname[1]

// в xPath работает примерно как .\* в регексах:

XPath 2.0 ▾ `//firstname`

zagoskin-blagorodnyj-teatr.xml ✘ teachers\_full.xml ✘ teachers\_with\_attrs.xml ✘

teachers person firstname

```
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3   <person gender="F">
4     <firstname>Нина</firstname>
5     <patronym>Евгеньевна</patronym>
6     <surname>Сахарова</surname>
7   </person>
8   <person gender="M">
9     <firstname>Борис</firstname>
10    <patronym>Валерьевич</patronym>
11    <surname>Орехов</surname>
12  </person>
13  <person gender="M" birthdate="1989-12-10">
14    <firstname>Даниил</firstname>
15    <patronym>Андреевич</patronym>
```

Text Grid Auth

Description - 6 items

Нина

Борис

Даниил

XPath location

/teachers[1]/person[1]/firstname[1]

/teachers[1]/person[2]/firstname[1]

/teachers[1]/person[3]/firstname[1]

Впрочем, звездочка \* тоже используется — это любое имя элемента:

The screenshot shows an XML editor interface with the following details:

- Toolbar:** Includes icons for file operations, search, and various tools.
- XPath Bar:** Displays the expression `//person/*`.
- Result List:** Shows three XML files:
  - `zagoskin-blagorodnyj-teatr.xml`
  - `teachers_full.xml*` (highlighted)
  - `teachers_with_attrs.xml`
- XML Content:** The `teachers_full.xml` content is displayed:

```
teachers person firstname
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3 <person gender="F" age="27" birthdate="1992-03-12">
4   <firstname>Нина</firstname>
5   <patronym>Евгеньевна</patronym>
6   <surname>Сахарова</surname>
7 </person>
8 <person gender="M" age="37" birthdate="1982-09-30">
9   <firstname>Борис</firstname>
10  <patronym>Валерьевич</patronym>
11  <surname>Орехов</surname>
12 </person>
13 <person gender="M" age="29" birthdate="1989-12-10">
14   <firstname>Даниил</firstname>
15   <patronym>Андреевич</patronym>
16   <surname>Скоринкин</surname>
17 </person>
```

# Запрос с атрибутом в xPath: @имяатрибута

XPath 2.0  //\*

zagoskin-blagorodnyj-teatr.xml teachers\_full.xml teachers\_with\_attrs.xml

teachers	person
1	<?xml version="1.0" encoding="UTF-8"?>
2	<teachers>
3	<person gender="F">
4	<firstname>Нина</firstname>
5	<patronym>Евгеньевна</patronym>
6	<surname>Сахарова</surname>
7	</person>
8	<person gender="M">
9	<firstname>Борис</firstname>
10	<patronym>Валерьевич</patronym>
11	<surname>Орехов</surname>
12	</person>
13	<person gender="M" birthdate="1989-12-10">
14	<firstname>Даниил</firstname>
15	<patronym>Андреевич</patronym>

Description - 6 items

F

M

M

XPath location

/teachers[1]/person[1]/@gender

/teachers[1]/person[2]/@gender

/teachers[1]/person[3]/@gender

Text

Grid

Auth

XPath 2.0 //person[@gender="F"]

zagoskin-blagorodnyj-teatr.xml • teachers\_full.xml • teachers\_with\_attrs.xml

```
teachers person
3 <person gender="F">
4   <firstname>Нина</firstname>
5   <patronym>Евгеньевна</patronym>
6   <surname>Сахарова</surname>
7 </person>
8 <person gender="M">
9   <firstname>Борис</firstname>
10  <patronym>Валерьевич</patronym>
11  <surname>Орехов</surname>
12 </person>
13 <person gender="M" birthdate="1989-12-10">
14   <firstname>Даниил</firstname>
15   <patronym>Андреевич</patronym>
16   <surname>Скоринкин</surname>
17 </person>
18 <person gender="M">
19   <firstname>Франк</firstname>
20   <patronym>Куртович</patronym>
21   <surname>Фишер</surname>
22 </person>
23 <person gender="M">
24   <firstname>Георгий</firstname>
25   <patronym>Алексеевич</patronym>
26   <surname>Мороз</surname>
27 </person>
28 <person gender="F">
29   <firstname>Ольга</firstname>
30   <patronym>Николаевна</patronym>
31   <surname>Ляшевская</surname>
32 </person>
```

Проверка  
значения  
атрибута у  
элемента

XPath 2.0 //person[@gender="M"]

zagoskin-blagorodnyj-teatr.xml teachers\_full.xml teachers\_with\_attrs.xml

	teachers	person
3	<person gender="F">	
4	<firstname>Нина</firstname>	
5	<patronym>Евгеньевна</patronym>	
6	<surname>Сахарова</surname>	
7	</person>	
8	<person gender="M">	
9	<firstname>Борис</firstname>	
10	<patronym>Валерьевич</patronym>	
11	<surname>Орехов</surname>	
12	</person>	
13	<person gender="M" birthdate="1989-12-10">	
14	<firstname>Даниил</firstname>	
15	<patronym>Андреевич</patronym>	
16	<surname>Скоринкин</surname>	
17	</person>	
18	<person gender="M">	
19	<firstname>Франк</firstname>	
20	<patronym>Куртович</patronym>	
21	<surname>Фишер</surname>	
22	</person>	
23	<person gender="M">	
24	<firstname>Георгий</firstname>	
25	<patronym>Алексеевич</patronym>	
26	<surname>Мороз</surname>	
27	</person>	
28	<person gender="F">	
29	<firstname>Ольга</firstname>	
30	<patronym>Николаевна</patronym>	
31	<surname>Ляшевская</surname>	
32	</person>	

Проверка  
значения  
атрибута у  
элемента

XPath 2.0 //person[@age<30]

zagoskin-blagorodnyj-teatr.xml \* teachers\_full.xml\* teachers\_with\_attrs.xml

```
teachers person firstname
1 <?xml version="1.0" encoding="UTF-8"?>
2 <teachers>
3 <person gender="F" age="27" birthdate="1992-03-12">
4   <firstname>Нина</firstname>
5   <patronym>Евгеньевна</patronym>
6   <surname>Сахарова</surname>
7 </person>
8 <person gender="M" age="37" birthdate="1982-09-30">
9   <firstname>Борис</firstname>
10  <patronym>Валерьевич</patronym>
11  <surname>Орехов</surname>
12 </person>
13 <person gender="M" age="29" birthdate="1989-12-10">
14   <firstname>Даниил</firstname>
15   <patronym>Андреевич</patronym>
16   <surname>Скоринкин</surname>
17 </person>
18 <person gender="M" age="42" birthdate="1977-07-12">
19   <firstname>Франк</firstname>
20   <patronym>Куртович</patronym>
21   <surname>Фишер</surname>
22 </person>
23 <person gender="M" age="29" birthdate="1990-08-05">
24   <firstname>Георгий</firstname>
25   <patronym>Алексеевич</patronym>
26   <surname>Мороз</surname>
27 </person>
28 <person gender="F" age="46" birthdate="1973-05-07">
29   <firstname>Ольга</firstname>
```

Проверка  
значения  
атрибута у  
элемента  
(больше/  
меньше)