TEI

Text Encoding Initiative

Обзор: Введение в TEI

Инициатива кодирования текста (TEI) представляет собой сообщество, занимающееся вопросами обработки текста в академической области цифровых гуманитарных наук, которое непрерывно работает с 1980-х годов.

Сообщество в настоящее время ведет список рассылки, собрания и серию конференций и поддерживает одноименный технический стандарт, журнал, вики, и другие инструменты.

Сфера применения TEI

Формат используется многими проектами по всему миру. Практически все проекты связаны с одним или несколькими университетами. Некоторые известные проекты, которые кодируют тексты с использованием TEI, включают:

Проект	Ссылка	Особенности
British National Corpus	http://www.natcorp.ox.ac.uk	100 million word snapshot of current English
Oxford Text Archive	http://ota.ox.ac.uk/	>1 GB of Linguistic data and electronic texts in 25 languages
Perseus Project	http://www.perseus.tufts.edu/	Greek and Latin texts
EpiDoc	http://epidoc.sourceforge.net/	Epigraphy and Papyrology
Women Writers Project	http://www.wwp.northeastern.edu/	Early modern women writers (Margaret Cavendish, Eliza Haywood, etc.)
New Zealand Electronic Text Centre	http://www.nzetc.org/	New Zealand and Pacific Islands texts
The SWORD Project	http://www.crosswire.org/sword/	Bible software, dictionaries, Christian literature
FreeDict	http://freedict.org	Bilingual dictionaries
Text Creation Partnership	http://www.lib.umich.edu/tcp/	Early English and American books
CELT	http://celt.ucc.ie/publishd.html	Ancient and Medieval Irish Manuscripts

Цели кодирования текста

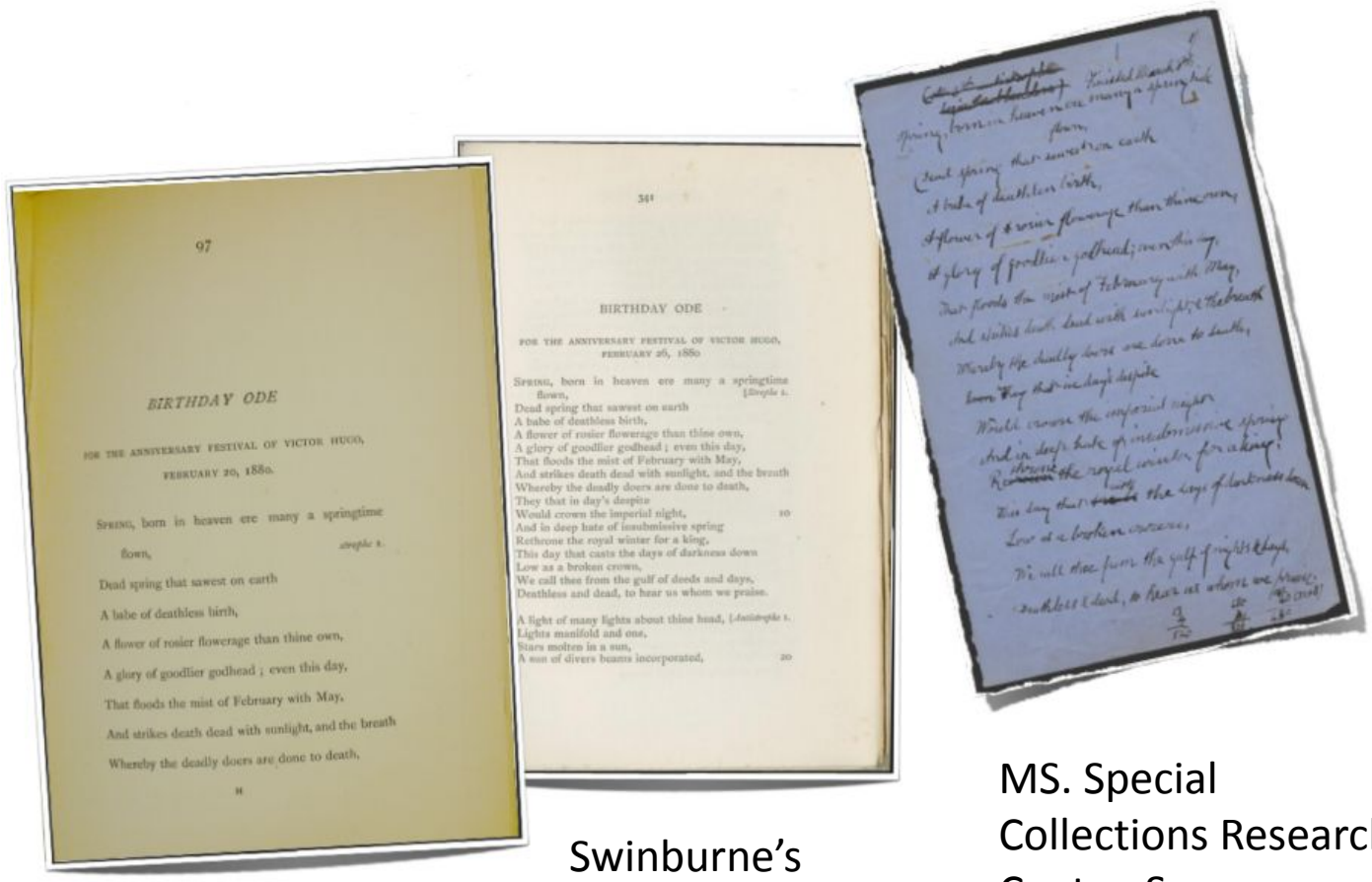
- Доступ и сохранение
- Распространение
  - Поиск / просмотр
  - Взаимодействие и переносимость между различными источниками
- Анализ
  - Лингвистический анализ
  - Тематическое моделирование
- Визуализация
  - Интерактивные временные рамки (см. VWWP)
  - Интерфейсы на основе карт (см. проект Swinburne)

Представление текста в кодировке

- Структурные особенности
- Текстовые разделы (главы, разделы и т.д.), абзацы, списки, таблицы, группы строк, строки и т. д.
- Контент и контекст:
  - Метаданные для электронного и исходного документа
  - Ссылки на людей, места, события, организации и т.д. в тексте (на уровене фраз)
  - Тематические и интерпретирующие аннотации
- Форматирование и дизайн
  - Полужирный шрифт, курсив, малый шрифт, подстрочный, цвет, размеры, привязки, водяные знаки и другие особенности исходного документа

Исходный текстовый документ

Варианты



97

BIRTHDAY ODE

FOR THE ANNIVERSARY FESTIVAL OF VICTOR HUGO,
FEBRUARY 26, 1880.

Spring, born in heaven ere many a springtime
flown, *strophe 1.
Dead spring that sawest on earth
A babe of deathless birth,
A flower of rosier flowerage than thine own,
A glory of goodlier godhead; even this day,
That floods the mist of February with May,
And strikes death dead with sunlight, and the breath
Whereby the deadly doers are done to death,

H

341

BIRTHDAY ODE

FOR THE ANNIVERSARY FESTIVAL OF VICTOR HUGO,
FEBRUARY 26, 1880

Spring, born in heaven ere many a springtime
flown, [Strophe 1.
Dead spring that sawest on earth
A babe of deathless birth,
A flower of rosier flowerage than thine own,
A glory of goodlier godhead; even this day,
That floods the mist of February with May,
And strikes death dead with sunlight, and the breath
Whereby the deadly doers are done to death,
They that in day's despite
Would crown the imperial night,
And in deep hate of insubmissive spring
Rethrone the royal winter for a king,
This day that casts the days of darkness down
Low as a broken crown,
We call thee from the gulf of deeds and days,
Deathless and dead, to hear us whom we praise.

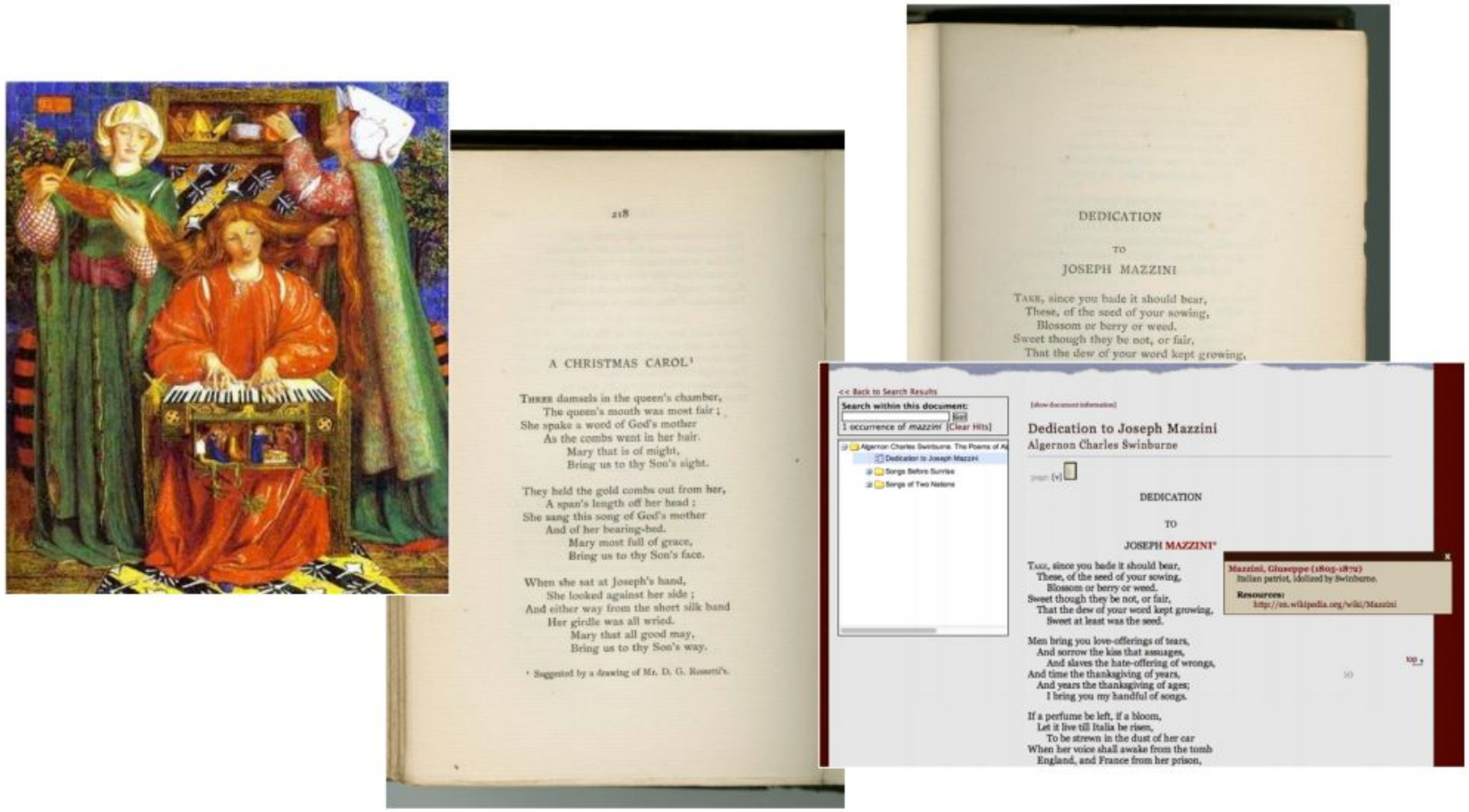
A light of many lights about thine head, [Antistrophe 1.
Lights manifold and one,
Stars molten in a sun,
A sun of divers beams incorporated,

Swinburne's Songs of the Springtides (1880)

Swinburne's Poems (1904)

MS. Special Collections Research Center. Syracuse University Library

Межтекстовая и контекстная информация



Преимущества кодирования текста

- Повторное использование и гибкость: создав один раз, можно использовать без ограничений
- Представление и вывод текста контролируется стилями (style sheets)
  - Можно создавать различные представления одного и того же текста и разных форматов: PDF, HTML, ePub (электронные книги), обычный текст (для текстового анализа) и т.д.
- Документ и разметка могут служить объектом анализа, причём поиск документов и информации в них упрощается

Особенности кодирования текста

- Текстовое кодирование не обязательно является простым вводом или распознаванием отсканированных документов; оно не объективно, а толковательно. Каждый закодированный текст является «чтением», интерпретацией исходного текста.
- Часто существует множество способов применения определенного языка разметки к определенному тексту.
- Каждый из проектов обычно требует рекомендаций и документации в дополнение к общей спецификации или рекомендациям по языку разметки.

TEI (Text Encoding Initiative)

- TEI:
  - официальная организация, Консорциум TEI;
  - научное сообщество - с ежегодной конференцией, изданием в открытом доступе и активным списком обсуждений по электронной почте.
  - стандарт кодирования текста, подготовленный этой организацией, Руководство TEI по кодированию и обмену электронными текстами.
- В наших целях TEI означает стандарт кодирования технического текста

История TEI

До создания TEI у ученых гуманитарных наук не было единых стандартов кодирования электронных текстов таким образом, который служил бы их академическим целям.

В 1987 году группа ученых, представляющих области гуманитарных наук, лингвистики и вычислительной техники, созванная в Колледже Вассара, представила ряд руководств, известных как «Принципы Покипси». Эти руководящие принципы направлены на разработку первого стандарта TEI, «P1».

- 1987 – началась работа над тем, что впоследствии станет называться TEI
- 1994 – выпущен стандарт TEI P3
- 2002 – выпущен стандарт TEI P4
- 2007 – выпущен стандарт TEI P5

Рекомендации TEI: Краткий обзор

- Инициатива кодирования текста (TEI) / Руководство по кодированию и обмену электронными текстами (TEI)
- «Руководящие принципы TEI» адресованы всем, кто работает с любым текстом в электронной форме, и предоставляют средства для представления тех функций текста, которые должны быть четко определены, чтобы облегчить обработку текста с помощью компьютерных программ
- TEI предлагает элементы, атрибуты и другие механизмы кодирования прозы, поэзии, драмы, словарей, и других научных и ненаучных текстов.

Рекомендации TEI: Краткий обзор

- Рекомендации TEI:
  - Могут применяться добуквенно или в вольной интерпретации
  - Разработаны как набор модулей / механизмов, которые могут быть выбраны по мере необходимости:
    - core: элементы, общие для всех документов TEI
    - figures: таблицы, рисунки, формулы, нотные обозначения
    - linking: ссылки, разбиение на абзацы, выравнивание
    - msdescription: описание рукописи
    - namesdates: имена и даты
  - Могут быть адаптированы под конкретные нужды

Рекоммендации TEI версии P5

- Рекоммендации к прозе с примерами: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/
- Набор элементов/тегов в версии P5: http://www.tei-c.org/release/doc/tei-p5-doc/en/html/REF-ELEMENTS.html (перечисление тегов с примерами и ссылками на документацию к прозе)

TEI P5: базовые компоненты

- <TEI>: корневой элемент документа TEI
  - <teiHeader>: заголовок метаданных для документа TEI. Включает библиографические, технические, административные и другие метаданные о цифровом файле и аналоговом источнике, если таковой существует.
  - <text>: сам текст, например титульная страница и главы романа, акты и сцены драмы, книги или песни большой поэмы. Элемент <text> далее подразделяется на:
    - <front>: фронт, например, титульная страница(ы), оглавление, возможно предисловие или посвящение
    - <body>: основная часть документа
    - <back>: окончание документа, например, индексы, приложения

TEI P5: Разметка прозы

- : (деление) используется для базовых структурных подразделений текста, например томов, глав, разделов, кантов, оглавлений, индексов, приложений и т.д. Атрибут «type» может использоваться для обозначения типа деления.
  - ... </ div>
  - ... </ div>
  - ... </ div>
  - ... </ div>
- <head>: (заголовок) содержит любой тип заголовка, например название раздела, или заголовок списка, рисунка, таблицы и т.д.
- <p>: (paragraph, абзац)
- <pb>: (page break, разрыв страницы) обозначает границу между одной страницей текста и следующей

TEI P5: Разметка прозы

Chapter 1: The Manor House

Charles hadn’t visited the manor house since Easter,1955, and now he remembered why. “Hullo”, he called out as he walked up the drive, and then, as if to himself, “To be or not to be?, to walk or not to walk...oh, hang it all!” His meditation on Hamlet was interrupted as he collided with a peacock. “Sacré bleu!” he exclaimed with irritation, his sang-froid completely deserting him. It was going to be a long week.

His catalog of irritations included:

1. The weather
2. The peacocks
3. His meager grasp of French

TEI P5: Разметка прозы

```
<?xml version="1.0" encoding="UTF-8"?>
<div type="chapter">
    <head>Chapter 1: The Manor House</head>
    <p>Charles hadn't visited the manor house since
        Easter,1955, and now he remembered why.</p>
    <p><said>Hullo</said>, he called out as he walked up the
        drive, and then, as if to himself, <said>To be or
         not to be?, to walk or not to walk...oh,
         <emph rendition="#b">hang</emph> it all!</said>
        His meditation on Hamlet was interrupted as he
        collided with a peacock. <said xml:lang="fr">Sacré
        bleu!</said> he exclaimed with irritation, his
        <foreign xml:lang="fr">sang-froid</foreign> completely deserting him.
        It was going to be a long week. His catalog of irritations included:
            <list type="ordered">
            <item>The weather</item>
            <item>The peacocks</item>
            <item>His meager grasp of French</item>
        </list>
    </p>
</div>
```

TEI P5: Разметка поэзии

- <lg>: (line group, группа строк) содержит группу стихотворных строк (стихов), функционирующих как формальная единица, например. строфа, рефрен, параграф стихотворения и т. д. Атрибуты type и subtype могут использоваться для классификации типа группы строк
- <l>: (line, строка) содержит строку стихотворения (стих)

TEI P5: Разметка поэзии

THE ROUNDEL

A ROUNDEL is wrought as a ring or a starbright sphere,
With craft of delight and with cunning of sound unsought,
That the heart of the hearer may smile if to pleasure his ear
A roundel is wrought.

Its jewel of music is carven of all or of aught—
Love, laughter, or mourning—remembrance of rapture or fear—
That fancy may fashion to hang in the ear of thought.

As a bird's quick song runs round, and the hearts in us hear
Pause answer to pause, and again the same strain caught,
So moves the device whence, round as a pearl or tear,
A roundel is wrought.

TEI P5: Разметка поэзии

```xml
<?xml version="1.0" encoding="UTF-8"?>
<div type="poem">
    <head rendition="#center #uppercase">The Roundel</head>
    <lg>
        <l><hi rendition="#small-caps">A roundel</hi> is wrought as a ring or a starbright sphere,</l>
        <l>With craft of delight and with cunning of sound unsought,</l>
        <l>That the heart of the hearer may smile if to pleasure his ear</l>
        <l rendition="#l-indent-03">A roundel is wrought.</l>
    </lg>
    <lg>
        <l>Its jewel of music is carven of all or of aught–</l>
        <l>Love, laughter, or mourning–remembrance of rapture or fear–</l>
        <l>That fancy may fashion to hang in the ear of thought.</l>
    </lg>
    <lg>
        <l>As a bird's quick song runs round, and the hearts in us hear</l>
        <l>Pause answer to pause, and again the same strain caught,</l>
        <l>So moves the device whence, round as a pearl or tear,</l>
        <l rendition="#l-indent-03">A roundel is wrought.</l>
    </lg>
</div>
```

TEI P5: Разметка драматургии

- <sp>: (speech, речь) содержит отдельную речь в тексте исполнения или отрывок, представленной в прозе или стиховом тексте.
- <speaker>: содержит специализированную форму заголовка или метки, дающую название одному или нескольким говорящим в драматическом тексте или фрагменте.
- <stage>: (описание сцены) содержит любое описание сцены в драматическом тексте или фрагменте.

TEI P5: Разметка драматургии

Scene 1

Enter Fay

Fay: I say, Dinah, has anyone seen my gloves?

Enter Dinah

Dinah:

No, miss, perhaps the parakeet has got them again?

Exit Fay and Dinah

TEI P5: Разметка драматургии

```xml
<?xml version="1.0" encoding="UTF-8"?>
<div type="scene">
    <head rendition="#center">Scene 1</head>
    <stage rendition="#i">Enter Fay</stage>
    <sp>
        <speaker>Fay:</speaker>
        <p>I say, Dinah, has anyone seen my gloves?</p>
    </sp>
    <stage rendition="#i">Enter Dinah</stage>
    <sp>
        <speaker>Dinah:</speaker>
        <p>No, miss, perhaps the parakeet has got them again?</p>
    </sp>
    <stage rendition="#i">Exit Fay and Dinah</stage>
</div>
```

TEI P5: Разметка писем

- <opener>: группирует строку с датой, с адресантом, приветствие и подобные фразы, представляя собой первую группу в начале разделения (div).
- <closer>: группирует строку с датой, с адресантом, прощание и подобные фразы, представляя собой окончательную группу в начале разделения (div).
  - <dateline>: содержит краткое описание места, даты, времени и т.д. написания письма, добавляемое к нему в начале или в конце
  - <salute>: (salutation, приветствие) содержит приветствие или прощание в конце письма, предисловия и т.д.
  - <signed>: (signature, подпись) содержит закрывающее прощание

TEI P5: Разметка писем

1906 August the 5th

Cape Cod

My dear Becky

How lovely the oysters are this evening!

Yours very truly

Maria

TEI P5: Разметка писем

```
<div type="letter">
    <opener>
        <dateline>
            <date when="1906-08-05">1906 August the 5th</date>
            <lb/>
            Cape Cod
        </dateline>
        <salute>
            My Dear Becky
        </salute>
    </opener>
    <p>How lovely the oysters are this evening!</p>
    <closer>
        <salute>Yours very truly</salute>
        <signed>Maria</signed>
    </closer>
</div>
```