Курс «Языки разметки»

Лекция 1. Введение в языки разметки

Содержание

- О курсе
- Понятие языка разметки
- История развития языков разметки

О курсе (ВМИ-111,112)

- Распределение часов:
  - Лекции – 18 час. (**пт. по 1-й неделе**)
  - Практические – 36 час.
- Форма контроля: зачет
- Страница курса: http://foreva.susu.ru/for_stud/ml/

О курсе (ВМИ-215)

- Распределение часов:
  - Лекции – 18 час. (**сб. по 1-й неделе**)
  - Практические – 18 час.
- Форма контроля: зачет
- Страница курса: http://foreva.susu.ru/for_stud/ml/

Категории информационных ресурсов

- **Данные** – сведения о сущностях предметной области, их свойствах и связях с другими сущностями.

Категории информационных ресурсов

- **Метаданные** – данные о данных, которые могут описывать не только свойства данных, но и свойства информационной системы в целом, ее отдельных механизмов и их функций, поддерживаемых технологий, пользователей и др.

- Основные назначения метаданных:
  1. Определить ***внешнее представление*** документа

Категории информационных ресурсов

Программа-преобразователь

Иванов И.И.

30 лет

Высшее, 2006, ЮУрГУ, магистр

Стаж 9 лет

- 2005-… Разноэкспорт, программист
- 2003-2004 МикроИТ, сист. администратор

Отличные знания английского языка

Категории информационных ресурсов

- **Метаданные** – данные о данных, которые могут описывать не только свойства данных, но и свойства информационной системы в целом, ее отдельных механизмов и их функций, поддерживаемых технологий, пользователей и др.

- Основные назначения метаданных:
  2. Осуществлять ***эффективный поиск***

Категории информационных ресурсов

+ метаданные с форматированием

Программа-преобразователь

Иванов И.И.

Возраст:
30 лет

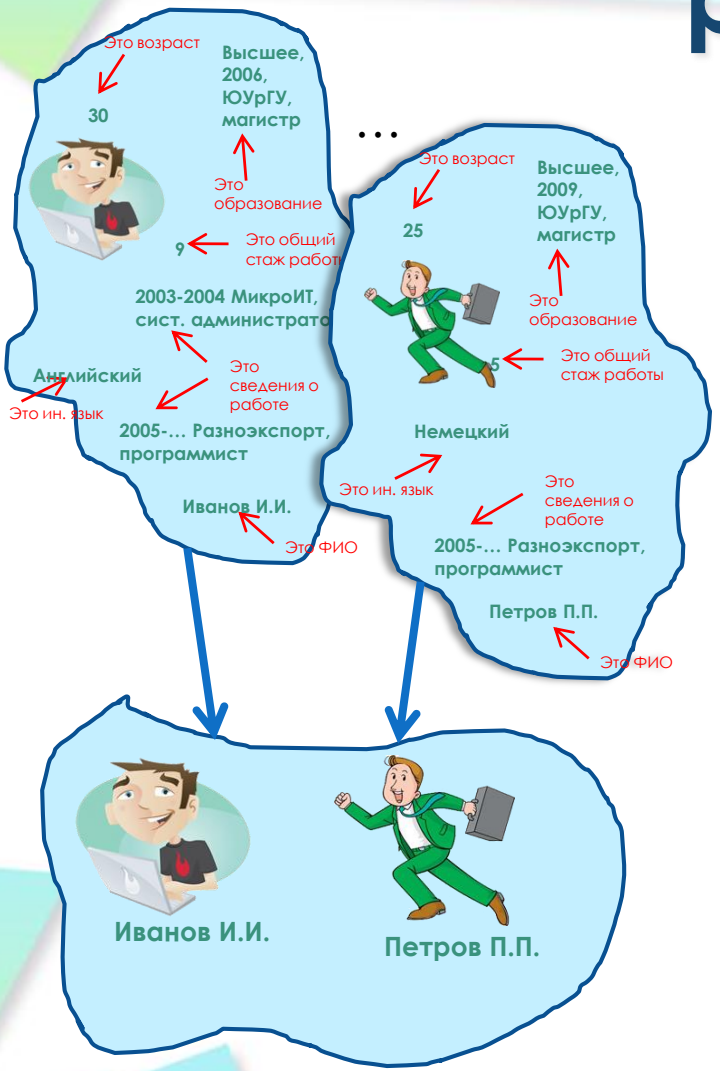
Образование:
Высшее,
2006, ЮУрГУ, магистр

Стаж работы: 9 лет

Опыт работы:
- 2005-… Разноэкспорт, программист
- 2003-2004 МикроИТ, сист. администратор

Знания языков: Английский

Категории информационных ресурсов



- **Метаданные** – данные о данных, которые могут описывать не только свойства данных, но и свойства информационной системы в целом, ее отдельных механизмов и их функций, поддерживаемых технологий, пользователей и др.

- Основные назначения метаданных:
  3. Обеспечить ***интеграцию информационных ресурсов***, повторно использовать как документ целиком, так и отдельные его элементы

Язык разметки

- **Язык разметки (markup language)** – средство описания данных и метаданных, хранящихся в документе.

Язык разметки $\neq$ *Язык программирования*

- Язык разметки ничего не вычисляет и не выводит данные на экран или принтер!!!

- Примеры:
  - Язык разметки гипертекста HTML
  - Язык разметки XML
  - Язык разметки векторной графики SVG
  - Язык разметки текста и формул T_EX

Основные элементы языка разметки

- Основными понятиями любого языка разметки являются *теги*, *элементы* и *атрибуты*.
- ***Теги (tags)*** – специальные символы, позволяющие отличать в документе описание разметки от описания данных.
- ***Элемент*** – это тэги в совокупности с их содержанием (данными).
- Примеры:
  - HTML: `<title>Резюме</title>`
  - TeX: `\title{Резюме}`

Основные элементы языка разметки

- Основными понятиями любого языка разметки являются *теги*, *элементы* и *атрибуты*.
- ***Теги (tags)*** – специальные символы, позволяющие отличать в документе описание разметки от описания данных.
- ***Элемент*** – это тэги в совокупности с их содержанием (данными).
- Примеры:
  - HTML: `<title>Резюме</title>`
  - TeX: `\title{Резюме}`

Основные элементы языка разметки

- ***Атрибут*** используется при определении элемента, чтобы задать какие-либо параметры, уточняющие характеристики данного элемента.
- Пример:
  - HTML: <h1 face="Arial Bold">Иванов И.И.</h1>

Виды разметки

- ***Стилистическая разметка*** отвечает за внешний вид документа. Пример разметки:

```
<font face="Arial Bold" size="16">Евгений
  Онегин</font>
<font face="Arial Bold" size="12"><i>А.С.
  Пушкин</i></font>
<font face="Times New Roman" size="12">
  В книгу вошел роман в стихах...</font>
```

Виды разметки

- ***Структурная разметка*** задает структуру документа.
- Пример структурной разметки:

```
<div>
  <h1>Евгений Онегин</h1>
  <h2>А.С. Пушкин</h2>
  <p>В книгу вошел роман в стихах...</p>
</div>
```

Виды разметки

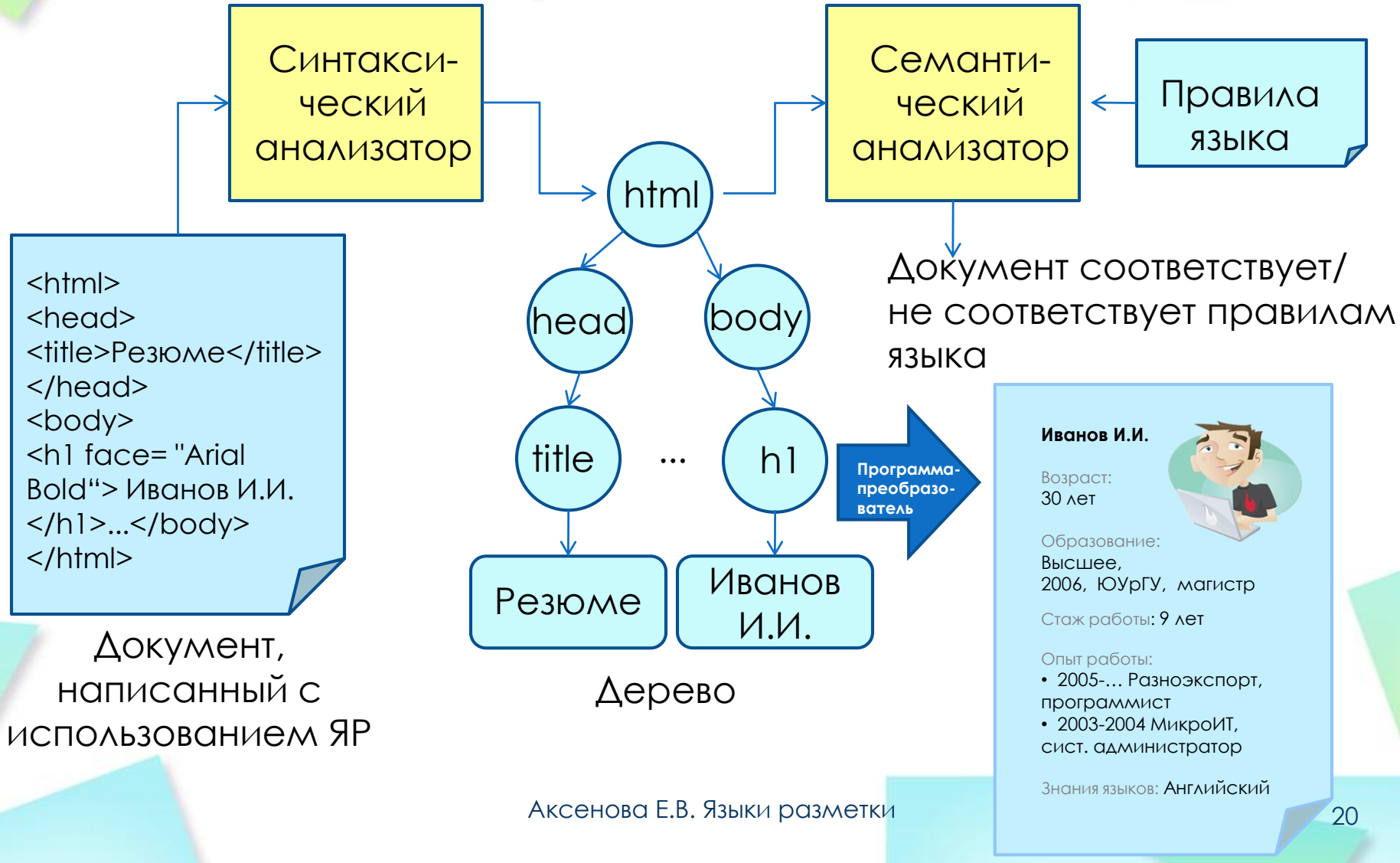
- ***Семантическая (контентная)*** разметка информирует о содержании данных.

```
<book>
  <title>Евгений Онегин</title>
  <author>А.С. Пушкин</author>
  <abstract>В книгу вошел роман в
  стихах...</abstract>
</book>
```

Категории языков разметки

- **Язык форматной разметки (Layout Markup или Presentation Markup)** обеспечивает возможность формирования представления размечаемых документов для их воспроизведения на мониторе компьютера или в твердой копии.
  - T_EX, RTF (Rich Text Format), PDF (Portable Document Format), Postscript, HTML (Hyper Text Markup Language).
- **Язык контентной разметки (Content Markup)** обеспечивает разметку, определяющую семантическую структуру документа.
  - SGML (Standard Generalized Markup Language), XML (eXtensible Markup Language).

Обработка документа с разметкой



Синтаксический анализ

- Программа производящая синтаксический анализ документа с разметкой называется ***парсером (parser).***
- Результат работы парсера – ***дерево документа.***

- Примеры синтаксических правил языка разметки XML:
  - Для каждого тега вида «<name_tag>» (открывающий тег) должен быть тег вида «</name_tag>» (закрывающий тег)
  - Теги <name> и <Name> - это разные теги.
  - Значение атрибута обязательно должно заключаться в кавычки.
- Все синтаксические правила записаны в **спецификации данного языка разметки**

для XML 1.0 спецификация опубликована по адресу http://www.w3.org/TR/REC-xml/

Семантический анализ

- ***Спецификация типа документа (Document Type Definition)*** – спецификация структуры документа, т.е. определение набора возможных разметок документов описываемого типа.
- Спецификация типа документа задает:
  - имена элементов, которые могут использоваться
  - порядок следования элементов
  - какие элементы являются обязательными, какие нет
  - и т.д.
- Соответствие спецификации типа документа определяют программы, называемые ***валидаторами (validator)***.

<!ELEMENT list - - (head?, item+)>	Список может иметь заголовок, список непуст.
<!ELEMENT head - 0 (#PCDATA)>	Заголовок списка – текст
<!ELEMENT item - 0 (p+)>	Элемент списка – непустой набор абзацев
<!ELEMENT p - 0 (#PCDATA)>	Абзац – текст.

Пример DTD для SGML

История развития языков разметки

- Термин «разметка» (markup) произошел от «marking up» (*помечание*, *размечание*) из традиционной издательской практики проставления специальных условных пометок на полях и в тексте рукописи или корректуры перед передачей ее в печать.
- Таким образом «разметчики» (markup men) указывали гарнитуру, стиль и размер шрифта для каждой части текста.

GenCode

- Идея использовать языки разметки в компьютерной обработке текстов была впервые обнародована Вильямом Тьюнниклиффом на конференции в 1967 году. Данная идея называлась «универсальное кодирование» («generic coding»).
- В 1970-е годы Тьюнниклифф руководил разработкой стандарта GenCode для издательской индустрии.

GML (Generalized Markup Language)

- Отцом языков разметки обычно называют исследователя IBM Чарльза Голдфарба (Charles Goldfarb).
  - Основная идея – 1969 г.
  - 1973 год - IBM GML
  - 1978 год – Голдфарб убедил руководство IBM использовать GML в коммерческих целях
  - 1980-е годы – разработка на базе GML и GenCode языка SGML

SGML

- **SGML** (*Standard Generalized Markup Language*) - **метаязык**, предназначен для создания на его основе специализированных языков разметки.
  - В 1980 г. – первый вариант спецификации.
  - В 1986 г. Международная организация по стандартизации ISO одобрила стандарт SGML ISO-8879.
- *Достоинства*: мощный метаязык разметки, позволяющий создавать языки разметки для различных предметных областей .
- *Недостатки*: большая сложность (по количеству, синтаксису и семантике объектов языка) затрудняет использование SGML в качестве языка разметки.
- Каждый язык разметки, определенный с помощью SGML, называется ***SGML приложением***. Например, HTML - SGML приложение.

SGML

- SGML-документ содержит следующие основные части:
  - ***Пролог (prolog):***
    - ***Объявление SGML (SGML declaration)***. Определяет, какие символы и разделители могут появляться в приложении.
    - ***Спецификация типа документов***.
  - ***Тело документа (document instance):***
    - ***Объекты документа***, содержащие данные и разметку. Каждый объект содержит ссылку на спецификацию типа документа, чтобы иметь возможность быть интерпретированным.

Объявление SGML

- Объявление SGML указывает основные данные об используемом диалекте SGML, такие, как набор символов, коды разделителей SGML, длину идентификаторов и т.д.
- Обычно объявление SGML существует в виде скомпилированных таблиц в SGML процессоре и пользователю не видимо.
- Пример:

```
<!SGML "ISO 8879:1986 (WWW)"
  --
  SGML Declaration для HyperText Markup Language версии HTML 4
  --
  CHARSET
      BASESET      "ISO Registration Number 177//CHARSET
                   ISO/IEC 10646-1:1993 UCS-4 with
                   implementation level 3//ESC 2/5 2/15 4/6«
  ...
>
```

Спецификация типа документов SGML

- <!DOCTYPE корневой_элемент PUBLIC/SYSTEM идентификатор_спецификации
  [здесь находятся все объявления для MY.DTD]
  >
- Пример:
  - <!DOCTYPE tei.2 SYSTEM "tei2.dtd" [...]>
- Пример правил языка для SGML-приложения HTML:
  - <!ELEMENT UL - - (LI)+>
  - <!ELEMENT IMG - O EMPTY>

Тело SGML

- Телом документа является собственно содержание документа. Оно содержит только текст, разметку и ссылки на обычные объекты, и, таким образом, не может содержать новых объявлений.

- Пример для SGML-приложения HTML:

```
<UL>
  <LI>Первый</LI>
  <LI>Второй</LI>
  <LI>Третий</LI>
</UL>
```

HTML

- Язык гипертекстовой разметки HTML (HyperText Markup Language) был предложен Тимом Бернерсом-Ли в 1989 году в качестве одного из компонентов технологии разработки распределенной гипертекстовой системы World Wide Web.
- Язык HTML является SGML-приложением.

HTML	<H1>Hi! What's up?</H1>
Результат	**Hi!** **What's up?**

XML

- XML (eXtensible Markup Language) — рекомендованный Консорциумом Всемирной паутины метаязык разметки.
- Язык XML – подмножество SGML. По сути является упрощением SGML.

XML	<fio>Иванов И.И.</fio> <age>30</age>

T_EX

Дональд Кнут
(р. 1938)

- Язык T_EX разработан в 1980-х Д. Кнутом для упрощения работы над книгой "Искусство программирования". Удобен для разработки научных документов (математика, физика и др.), поскольку имеет богатый инструментарий представления формул.

T_EX	\left\{ {\varphi ^t (x_0)} \right\}_{t = 1}^{ +\infty } \mathrel{\mathop{\kern0pt\longrightarrow} \limits_{}} \,\,\,\overline x \, \in \,\,\bigcap\limits_j {P_j \,\, = \,\,M}
Формула	$\left\{\varphi^t(x_0)\right\}_{t=1}^{+\infty} \longrightarrow \overline{x} \in \bigcap\limits_j P_j = M$

PDF и RTF

- Форматы *PDF – Portable Document Format* (компания Adobe Systems) и *RTF – Rich Text Format* (компания Microsoft) разработаны для обеспечения мобильности страничных документов, содержащих текст и графику. Документы в данных форматах содержат информацию, необходимую для отображения используемых шрифтов и др.

RTF	{\rtf1\ansi{\fonttbl\f0\fswiss Arial;}\f0 Hi!\par {\b What's up?} \par }
Результат	Hi! **What's up?**

Языки стилей

- ***Язык DSSSL (Document-Style Semantics and Specification Language)*** – язык управления способом форматирования SGML-документов для отображения их web-обозревателями и др. прикладными программами на базе механизма таблиц стилей.
- ***Язык CSS (Cascading-Style Sheets)*** – язык управления способом форматирования HTML-документов.
- ***Расширяемый язык таблиц стилей XSL (eXtensible Stylesheet Language)*** – язык управления способом форматирования XML-документов. В отличие от языка CSS, разработан специально для использования в среде XML и использует синтаксис XML.
- ***Язык XSL***, помимо обеспечения форматирования XML-документов, позволяет описывать трансформацию XML-документа в документ с другой разметкой и форматированием (например, трансформация XML-документа в HTML-документ).

Язык разметки HTML (Hyper Text Markup Language)

- HTML имеет теги для всех типов разметки. Но преимущественно и изначально ориентирован на стилистическую разметку.
- Пример стилистической разметки:

```
<font face="Arial Bold" size="16">Евгений Онегин</font>
```

- Пример структурной разметки:

```
<div>
  <h1>Евгений Онегин</h1>
  <h2>А.С. Пушкин</h2>
  <p>В книгу вошел роман в стихах...</p>
</div>
```

- Примерами семантического типа разметки в HTML являются теги <TITLE> </TITLE>(имя документа), <CODE> </CODE>(код, используется для листингов кода), <VAR> </VAR>(переменная), <ADDRESS> </ADDRESS>(адрес автора).

Структура HTML-документа

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
   "http://www.w3.org/TR/html4/strict.dtd">
<HTML>
    <HEAD>
        <TITLE>My first HTML document</TITLE>
    </HEAD>
    <BODY>
        <P>Hello world!
    </BODY>
</HTML>
```

SGML

- SGML-документ содержит следующие основные части:
  - ***Пролог (prolog):***
    - ***Объявление SGML (SGML declaration)***. Определяет, какие символы и разделители могут появляться в приложении.
    - ***Спецификация типа документов***.
  - ***Тело документа (document instance):***
    - ***Объекты документа***, содержащие данные и разметку. Каждый объект содержит ссылку на спецификацию типа документа, чтобы иметь возможность быть интерпретированным.

Структура HTML-документа

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
   "http://www.w3.org/TR/html4/strict.dtd">
```

Спецификация типа документа

```
<HTML>
   <HEAD>
      <TITLE>My first HTML document</TITLE>
   </HEAD>
   <BODY>
      <P>Hello world!
   </BODY>
</HTML>
```

Объекты документа

Спецификация типа документа HTML

- RFC 1866 — HTML 2.0, одобренный как стандарт 22 сентября 1995 года;
- HTML 3.2 — 14 января 1997 года;
- HTML 4.0 — 18 декабря 1997 года;
- HTML 4.01 — 24 декабря 1999 года;
- ISO/IEC 15445:2000 (так называемый ISO HTML, основан на HTML 4.01 Strict) — 15 мая 2000 года.
- HTML 5 — в разработке

```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01//EN"
"http://www.w3.org/TR/html4/strict.dtd">
```

Обработка HTML-документов

- HTML интерпретируется браузером, который производит и разбор и визуализацию.
- Валидация HTML-кода: http://validator.w3.org/
- Для форматирования элементов HTML используется язык каскадных стилей CSS.