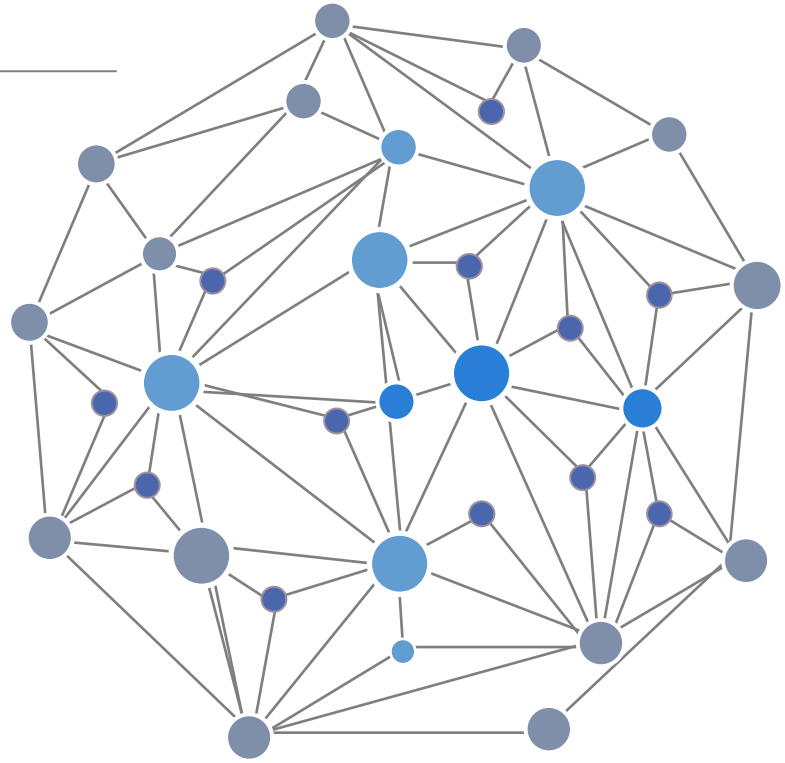
Web 程序设计



第三讲 XML 标准基础

福州大学 计算机与大数据学院
软件工程系 陈昱
2021-11

什么是 Web 标准?

Web 标准是一些规范的集合，是由 **W3C** 和其他的标准化组织共同制定的，用以创建和解释基于 Web 的内容的**规范**

结构化语言
* HTML （超文本置标语言） 4.01
* XHTML （可扩展超文本置标语言） 1.0
* XHTML 1.1
* XML （可扩展置标语言） 1.0 1.1

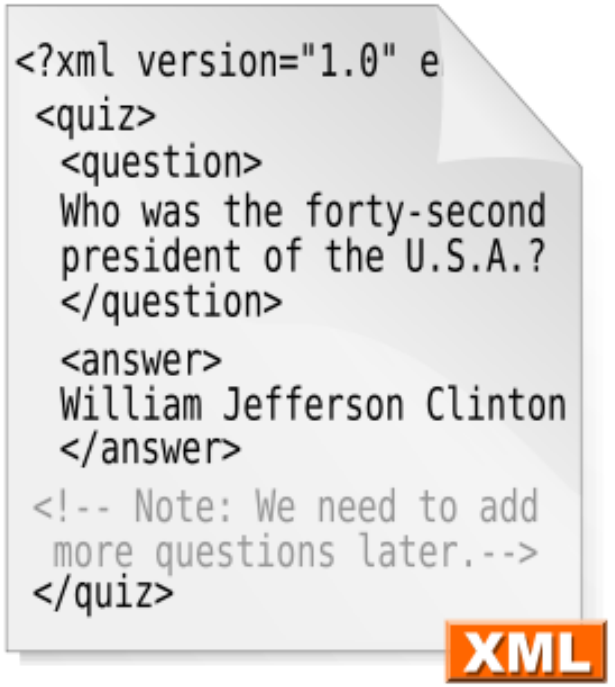
表现类语言
* CSS （层叠式样式表） Level 1
* CSS Level 2 revision 1
* CSS Level 3 （正在开发中）
* MathML （数学置标语言）
* SVG （矢量图形语言）

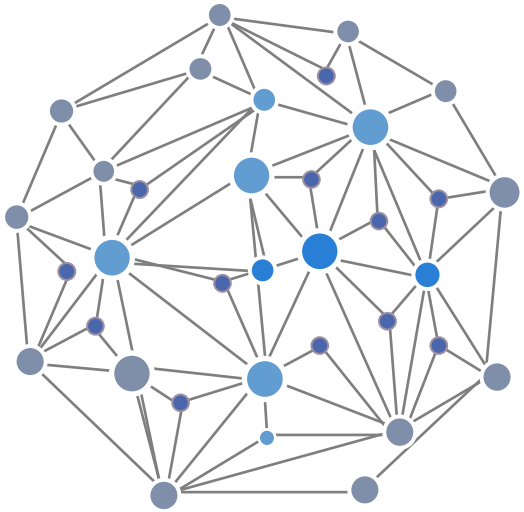
对象模型
* DOM （文档对象模型） Level 1
* DOM Level 2
* DOM Level 3 Core

脚本语言
* ECMAScript 262 （JavaScript的标准化版本）

XML Fundamental

- 置标语言
- 什么是 XML
- XML 文件的组成
- 文档类型定义 DTD
- XML 命名空间 Namespace
- XML 与 HTML 的比较





置标语言

Markup Language

置标语言

- 置标语言，一种用来给文本**添加标记**的语言

- 比如 HTML，它描述了一系列标记，每个标记表明了一定的信息（语义或是表现的）

Sample 1

- 标记 的含义是要求 HTML 浏览器这段文本表示一个段落
- 标记的含义是告诉浏览器这段文本需要强调

- Hello,World

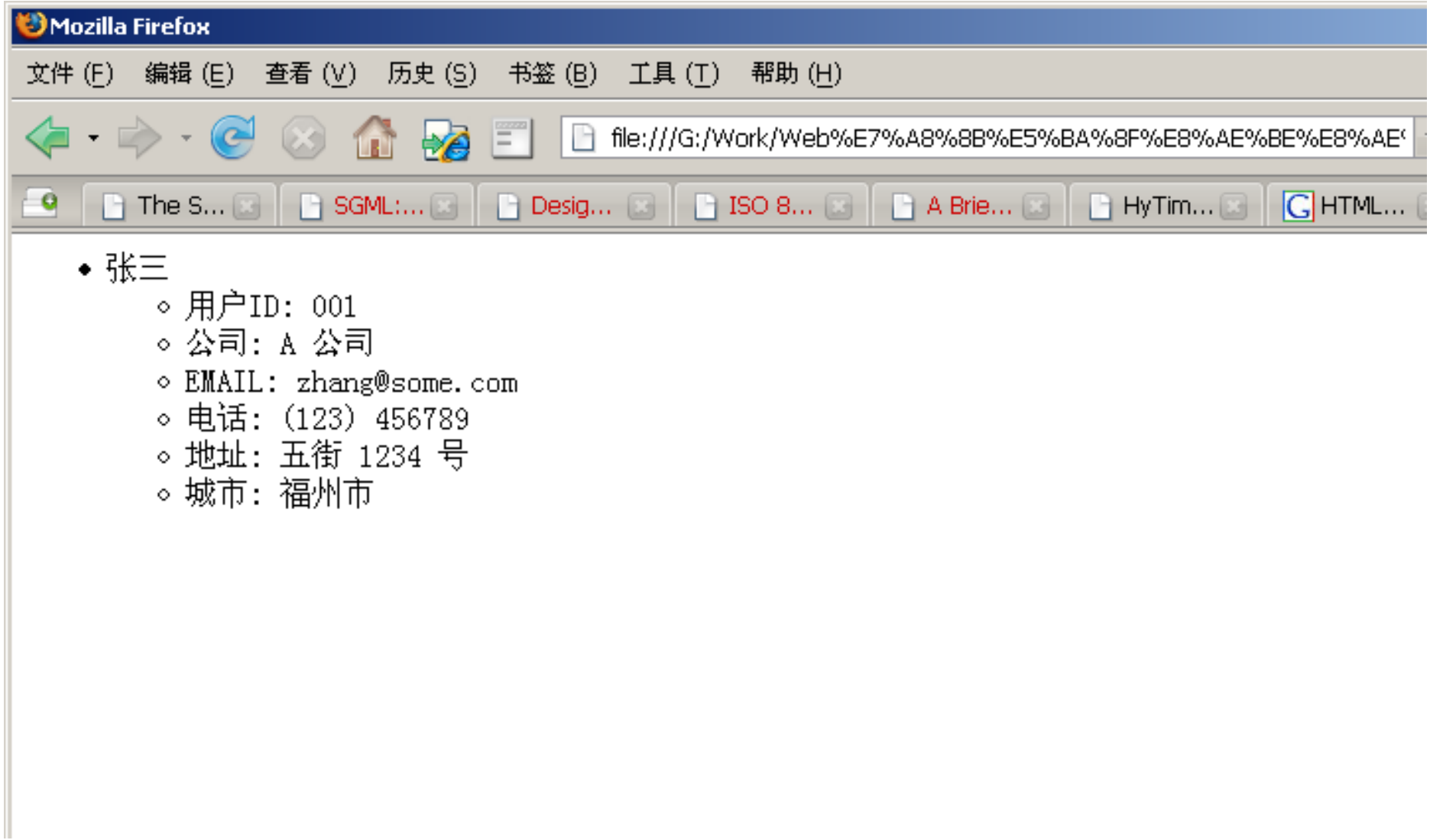
Hello,Wolrd

Sample 2

下面这一段 HTML 代码显示了一个客户联系信息列表：

```
<ul>
<li>张三</li>
    <ul>
        <li>用户ID: 001</li>
        <li>公司: A 公司</li>
        <li>EMAIL: zhang@some.com</li>
        <li>电话: (123) 456789</li>
        <li>地址: 五街 1234 号</li>
        <li>城市: 福州市</li>
    </ul>
</ul>
```

Sample 2


Mozilla Firefox
文件 (F) 编辑 (E) 查看 (V) 历史 (S) 书签 (B) 工具 (T) 帮助 (H)
file:///G:/Work/Web%E7%A8%8B%E5%BA%8F%E8%AE%BE%E8%AE
The S...
SGML:...
Desig...
ISO 8...
A Brie...
HyTim...
HTML...
张三
用户ID: 001
公司: A 公司
EMAIL: zhang@some.com
电话: (123) 456789
地址: 五街 1234 号
城市: 福州市

“置标” Markup

- “置标”的精确定义是：
  就数据本身的信息对数据进行编码的方法

- “置标” 的概念在现实生活中很常见
  - 比如用笔在课本上划重点

置标语言的定义方式

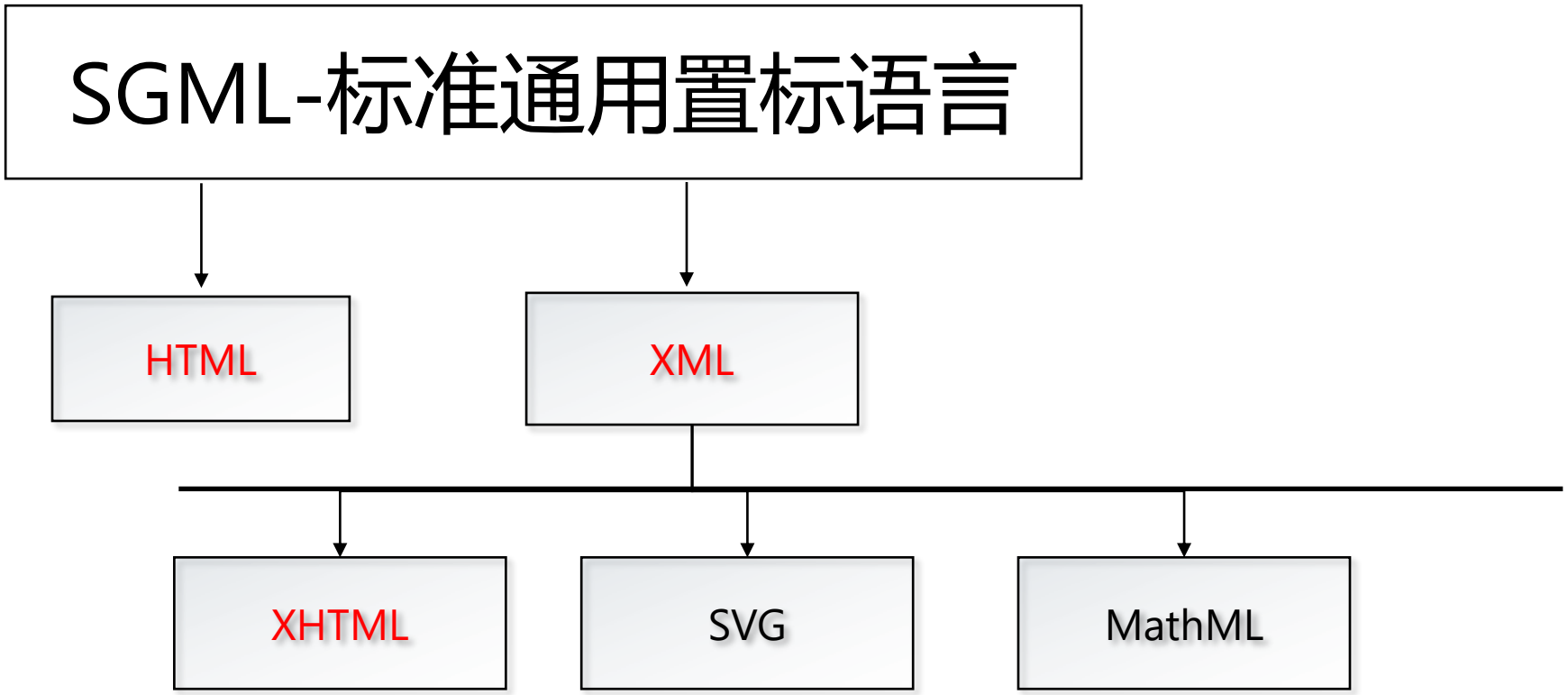
- 当我们需要通过标记将有用的信息告知人或是计算机时，需要定义两件事情：

- 首先，我们必须有一个标准，用它来描述什么是标记 **(在HTML中为<tag>)**

- 其次，我们还要有一个标准描述每个标记的具体含义 **(在HTML规范中定义)**

置标语言家族一览

- **SGML**, 鼻祖, 元语言 (Meta Language)
- **HTML**, 第一个 **Web** 置标语言

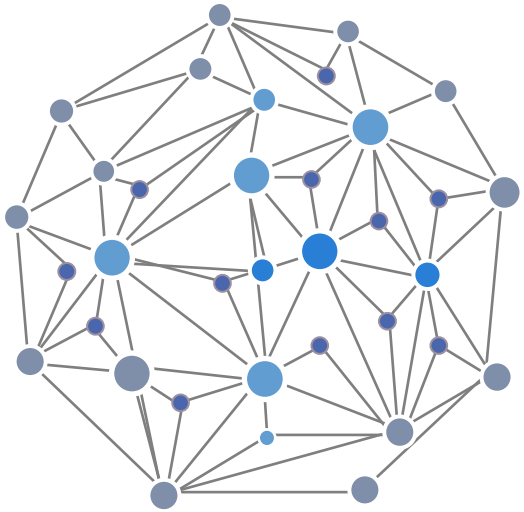


HTML 演化

- HTML 1 (Berners-Lee, 1990): 非常简单，有限的多媒体
  - in 1993, Mosaic 浏览器添加了一些新特性 (比如图像)
- HTML 2.0 (IETF, 1995): 试图标准化这些新特性, 但是 ...
  - 1995-97, Netscape & IE 浏览器大战
- HTML 3.2 (W3C, 1997): 尝试制定统一的标准
  - 但跟不上一些新技术，如 Java Applets & 流媒体
- HTML 4.01 (W3C, 1999): 后来十五年的正式标准
- HTML 5 (W3C & WHATWG): 已正式发布

早期 HTML 的局限性

1. 逐渐成为描述信息显示的工具
   - 数据搜索不易，搜索引擎无法理解
2. 浏览器大战导致浏览器兼容性
   - 标记越来越多
3. 可扩展性差
   - 无法适应各行各业的特殊要求（数学，化学）
4. 书写时缺乏严格的语法检查机制
   - 内部条理性差



XML

可扩展置标语言

XML

- XML 是 eXtensible Markup Language (可扩展置标语言) 的简写

- 和 HTML 一样，XML 同样来源于 SGML，但 XML 也是一种能定义其他语言的语言
  - XML 是 SGML 的子集
  - XML 也是元语言

XML 的版本

- 1998年2月10日 XML 1.0
  - 2008年11月26日 XML 1.0 (Fifth Edition)
- 2004年2月4日 XML 1.1
  - 2006年8月16日 XML 1.1 (Second Edition)

- 目前推荐使用的是 **XML 1.0**
  - 如果不需要 1.1 的新特性的话
- 参见 http://www.w3.org/TR/xml/

XML 的最大优势 - eXtensible

- 回顾刚才的 HTML 客户信息列表

- 尽管这也是一个存储、显示数据的可行的方法，它的效率和能力却非常有限
  - 数据的显示方式被固定了
  - 在这些数据中寻找信息困难
  - 提取数据困难，HTML 标记对数据理解无帮助

Sample: 客户联系信息列表

下面这一段 HTML 代码显示了一个客户联系信息列表：

```
<ul>
<li>张三</li>
    <ul>
        <li>用户ID: 001</li>
        <li>公司: A 公司</li>
        <li>EMAIL: zhang@some.com</li>
        <li>电话: (123) 456789</li>
        <li>地址: 五街 1234 号</li>
        <li>城市: 福州市</li>
    </ul>
</ul>
```

XML Solution

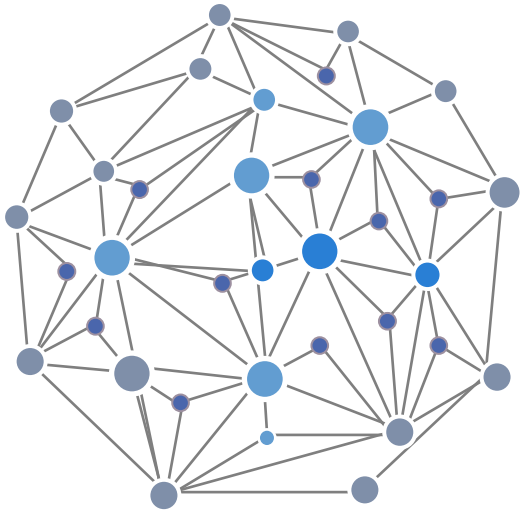
```
<联系人列表>
        <联系人>
                <姓名>张三</姓名>
                <ID>001</ID>
                <公司>A 公司</公司>
                <EMAIL>zhang@some.com</EMAIL>
                <电话>(1234) 456789</电话>
                <地址>
                        <街道>五街 1234 号</街道>
                        <城市>福州市</城市>
                        <省份>福建</省份>
                        <ZIP>350001</ZIP>
                </地址>
        </联系人>
</联系人列表>
```

XML 的优点

- 自我描述语言
- 便于人类和机器理解
- 遵循严格的语法要求

- 与 HTML 不同，XML 让你可以定义自己的标记!

（Extensible！）



XML 文件的组成

XML 文件的组成

- XML 声明
- 元素
- 属性
- 实体
- CDATA
- 注释

```
<?xml version = "1.0"?>
<小纸条>
  <收件人>张三</收件人>
  <发件人>李四</发件人>
  <主题>问候</主题>
  <具体内容>最近可好? </具体内容>
</小纸条>
```

XML 声明

- 一个最简单的 XML 声明是这样的：

  <?xml version="1.0"?>

- 声明中还有两个**可选**属性，分别是“standalone” 和 “encoding”

```
<?xml version = "1.0"
        standalone = "yes"
        encoding = "UTF-8"?>
```

XML 元素

- 元素是 XML 文件内容的基本单元
- 一个元素包含一个起始标记、一个结束标记以及标记之间的数据内容

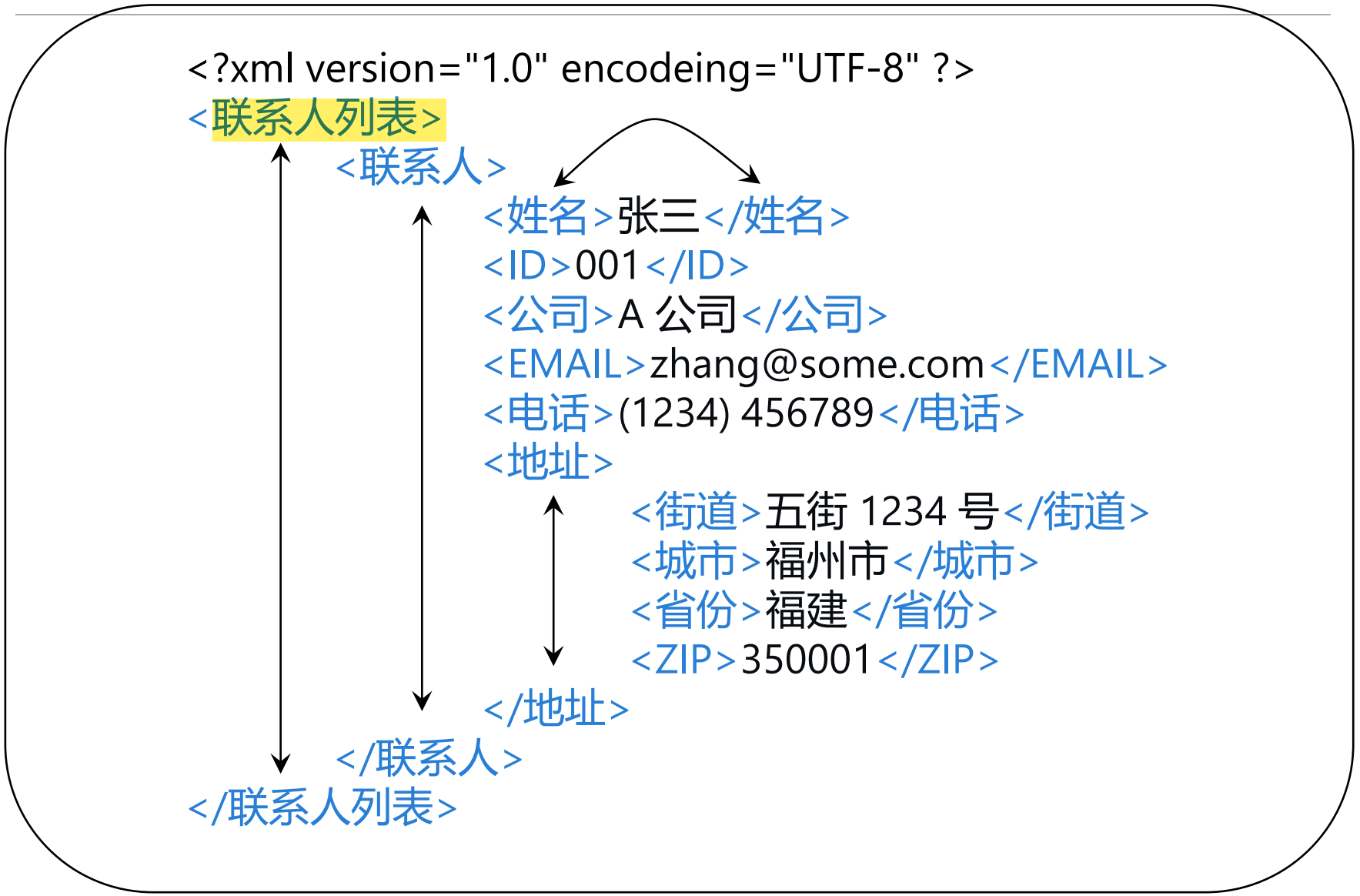
- 其形式是：<标记>数据内容</标记>
- 例如：　<姓名>张三</姓名>

标记

- 所有“<”和“>”之间的内容都称为标记

- 要求：
  - 标记名必不可少
  - 大小写有所区分
  - 要有正确的结束标记 <tag></tag>
  - 标记要正确嵌套 <a>

XML 根元素/正确嵌套


<?xml version="1.0" encodeing="UTF-8" ?>
<联系人列表>
<联系人>
<姓名>张三</姓名>
<ID>001</ID>
<公司>A 公司</公司>
<EMAIL>zhang@some.com</EMAIL>
<电话>(1234) 456789</电话>
<地址>
<街道>五街 1234 号</街道>
<城市>福州市</城市>
<省份>福建</省份>
<ZIP>350001</ZIP>
</地址>
</联系人>
</联系人列表>

合法的标记命名 lexical rule

- 以字母、下划线“_”或冒号“：”开头
- 后面跟随字母、数字、句号“.”、冒号、下划线或连字符“-”，但是中间不能有空格
- 而且任何标记名不能以“xml”开头！

- 另外，最好不要在标记的开头使用**冒号**，尽管它是合法的，但可能会带来混淆

属性 attribute

- 标记可以拥有属性，使得标记在描述信息时更加灵活　<标记名 （属性名="属性取值"）* >

<圆柱体 半径="10" 高="13">

- 注意：
  - 属性必须用**双引号**标记起来
  - 属性的值都被 XML 处理程序看作是字符串处理

字符数据

- 在 XML 中，起始和结束标记之间出现的所有合法字符都被忠实地传给 XML 处理程序

(1)

```
<格式>一段文字</格式>
```

(2)

```
<格式>
一段文字
</格式>
```

实体引用

- 为了避免把字符数据和标记中需要用到的一些特殊符号相混淆，XML 还提供了一些有用的实体引用替代数据中的特殊符号

字符	实体引用
<	&lt;
>	&gt;
&	&amp;
"	&quot;

实体引用

- 如果我们需要在“示例”这个标记中出现文本

<姓名>张三</姓名>

- 正确的写法是：

<示例><姓名>张三</姓名></示例>

CDATA

- 在一个特殊的标记 CDATA 下，所有的标记、实体引用都被忽略，而被 XML 处理程序一视同仁地当作字符数据看待

- CDATA 的形式如下：

```
<![CDATA[
        文本内容
]]>
```

CDATA Sample

```
<示例>
  <![CDATA[
      <联系人>
            <姓名>张三</姓名>
            <EMAIL>zhang@s.com</EMAIL>
      </联系人>
  ]]>
</示例>
```

注释

- XML 中注释是用“<!--”和“-->”引起来的字符串

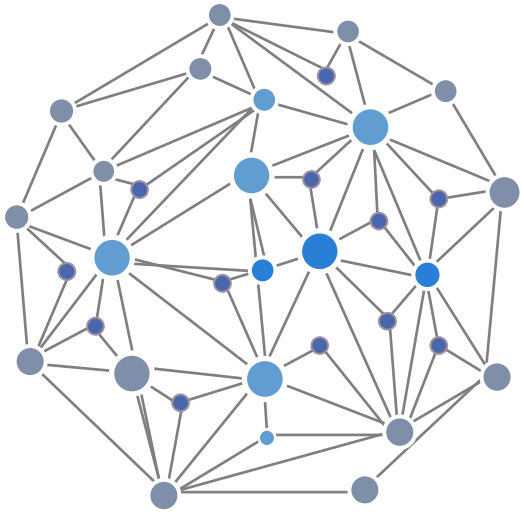
<示例>

<!-- 一个XML 的例子 -->

……

注释的注意事项

- 在注释文本中不能出现字符“-”或字符串“--”
- 不要把注释文本放在标记之中
- 不要把注释文本放在实体声明中，也不要放在 XML 声明之前
- 注释不能被嵌套



文档类型定义 DTD

Document Type Definition

形式良好 (Well-formed) XML 文件

- 所谓 “**形式良好**” 有着明确的标准，就是要**遵守 XML 规范中的语法规则**
  - 所有的标记都要正确的嵌套
  - 只有一个根元素
  - …
- 如何检查是否形式良好?
  - 用浏览器 (IE, Firefox etc.)
  - Eclipse, Visual Studio 等工具

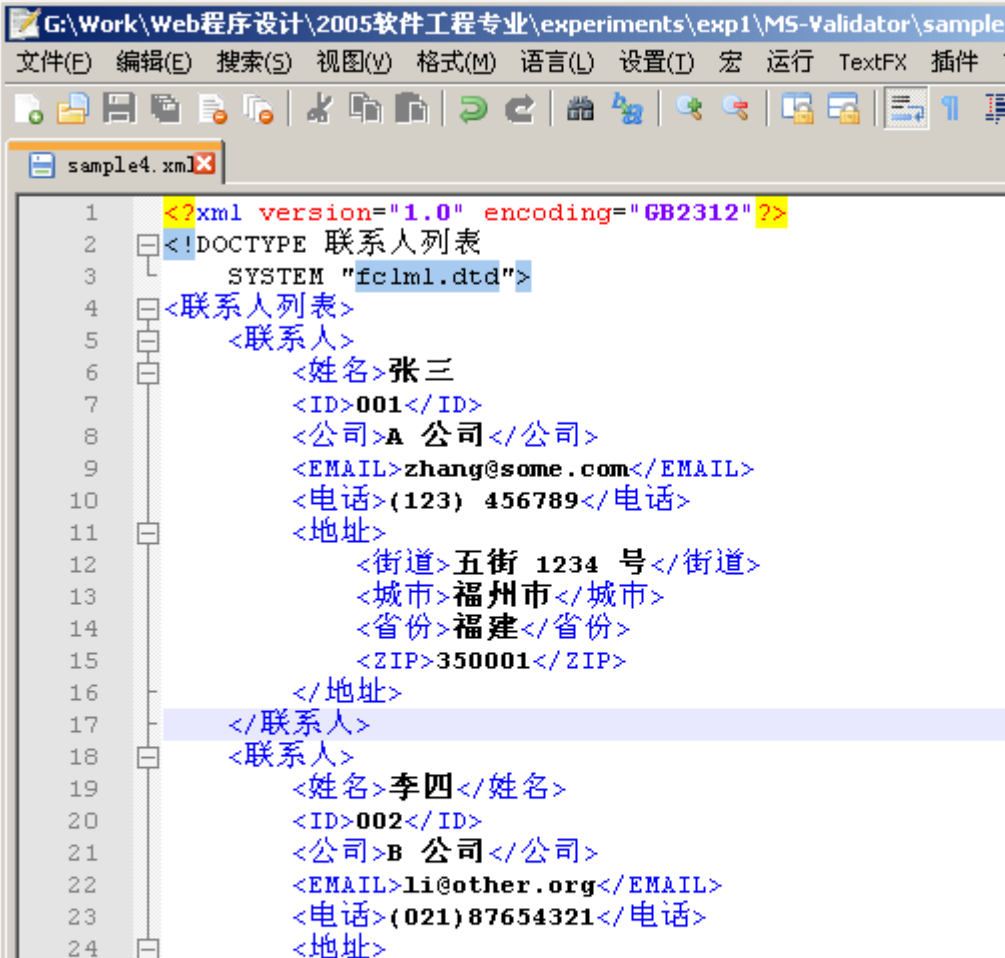
如何检查是否形式良好?

无法显示 XML 页。

使用 XSL 样式表无法查看 XML 输入。请更正错误然后单击 刷新按钮，或以后重试。

结束标记 '联系人' 与开始标记 '姓名' 不匹配。处理资源 'file:///G:/Work/Web程序设计/2005软件工程专业/experiments/exp1/MS-Validator/sample4.xml' 时出错。第 17 行...

```
 </联系人>
---^
```




```
1  <?xml version="1.0" encoding="GB2312"?>
2  <!DOCTYPE 联系人列表
3      SYSTEM "fclml.dtd">
4  <联系人列表>
5      <联系人>
6          <姓名>张三
7          <ID>001</ID>
8          <公司>A 公司</公司>
9          <EMAIL>zhang@some.com</EMAIL>
10         <电话>(123) 456789</电话>
11         <地址>
12             <街道>五街 1234 号</街道>
13             <城市>福州市</城市>
14             <省份>福建</省份>
15             <ZIP>350001</ZIP>
16         </地址>
17     </联系人>
18     <联系人>
19         <姓名>李四</姓名>
20         <ID>002</ID>
21         <公司>B 公司</公司>
22         <EMAIL>li@other.org</EMAIL>
23         <电话>(021)87654321</电话>
24         <地址>
```

有问题的例子

```
<?xml version="1.0" encoding="UTF-8"?>

<联系人>
  <联系人列表>
      <姓名>张三</姓名>
      <公司>A 公司</公司>
  </联系人列表>
    <电话>(123)456789</电话>
</联系人>
```

“有效”(Valid) 的 XML 文件

- 一个 XML 文件除了应该是 “形式良好”的外，还应该是“**有效**”使用了自定义标记的 XML 文件

- 所谓 “有效” 的 XML 文件也就是你的自定义标记必须遵循特定的使用规则（可以是你制定的，也可以是国际组织制定的）

文档类型定义 DTD

- DTD (Document Type Definition) 描述了一个置标**语言的语法和词汇表**，也就是定义了文件的整体结构以及文件的语法

```
<?xml version = "1.0" encoding="UTF-8"
      standalone = "yes"?>
<!DOCTYPE 根元素名[
    元素描述
]>
文件体.....
```

DTD Sample

```
<?xml version="1.0" encoding="UTF-8"?>
<!DOCTYPE 联系人列表 [
   <!ELEMENT 联系人列表 (联系人)*>
   <!ELEMENT 联系人 (姓名, 公司, 电话)>
   <!ELEMENT 姓名 (#PCDATA)>
   <!ELEMENT 公司 (#PCDATA)>
   <!ELEMENT 电话 (#PCDATA)>
]>
<联系人列表>
   <联系人>
      <姓名>张三</姓名>
      <公司>A 公司</公司>
      <电话>(123)456789</电话>
   </联系人>
</联系人列表>
```

DTD

外部 DTD

- 可以将 DTD 独立成一个文件，供多个 XML 引用
- 外部 DTD 常用于引用作者自己编写的 DTD

```
<!DOCTYPE 根元素名
        SYSTEM "外部 DTD 文件的URL">
```

公用 DTD

- 还存在一种外部 DTD，它是一个由权威机构制订的，提供给特定行业或公众使用的 DTD

- 因此，另一个引用外部 DTD 的办法是使用关键字 PUBLIC，引用这一类公开给公众使用的 DTD

公用 DTD

- 引用公共DTD 的形式为：

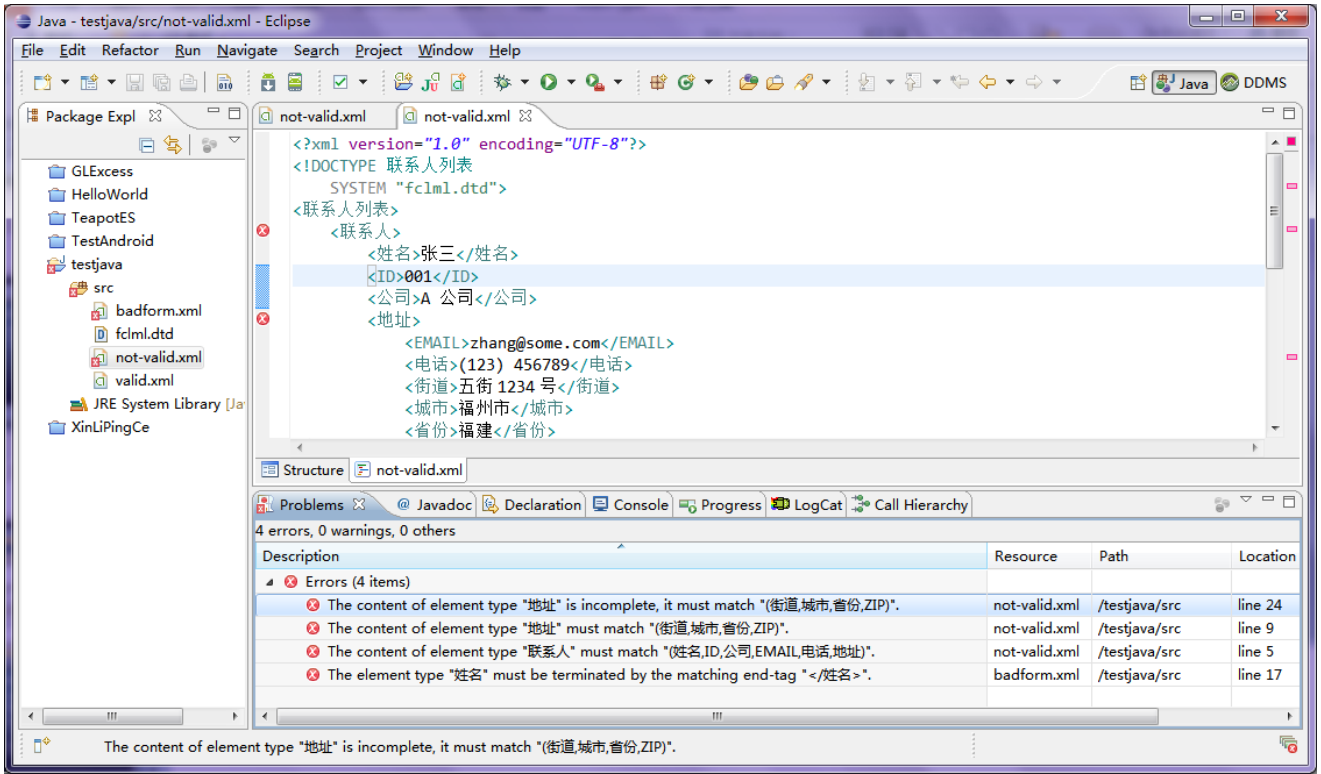
```
<!DOCTYPE 根元素
        PUBLIC "DTD 名称" "公用 DTD 的URL">
```

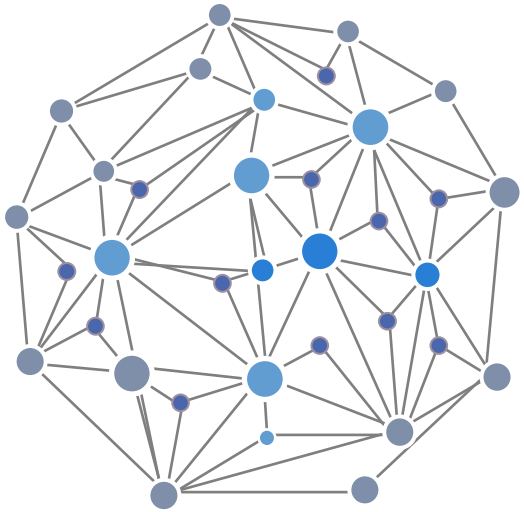
- 例如：

```
<!DOCTYPE 联系人列表 PUBLIC "联系人DTD"
        "http://www.domain.com/dtds/clml.dtd">
```

如何检查有效性?

- 利用XML验证工具检查 (XML validation tools)，有很多这类工具
  - Eclipse, NetBeans 等都带有这个功能





XML 命名空间

XML 命名空间 (Namespace)

- 如何在一个 XML 文档中，包含由多个DTD描述的元素?
- 不同 DTD 中可能定义了相同的标记名称，但是含义却是不同的
- 解决方法就是使用命名空间

XML 命名空间

- XML 解决方案：使用“唯一名称:元素名称”的形式
- 这个“唯一名称”就称为命名空间
- 技巧：使用 URI 来获得一个唯一的名称

- 例如：福州大学定义了一个 **student** 标记：**http://www.fzu.edu.cn:student**

XML 命名空间

```
<student xmlns:FZUST="http://fzu.edu.cn">
  <FZUST:stuno>220200xxx</FZUST:stuno>
  <FZUST:name>张三</FZUST:name>
  <FZUST:age>20</FZUST:age>
</student>
```

为何用 XML 做网页没有普及?

- 既然 XML 这么好，为何我们很少见过用 XML 做的网页？？？
- 优点有很多，但是一套自定义标记就是一种语言，开发效率低下，开发代码复杂

下一讲预告

解决之道：
XHTML & HTML5

用 XML 重新定义HTML

第三讲课后练习

- 学习
  - https://www.w3school.com.cn/xml/
- 辅助阅读
  - XML中国论坛《XML初学进阶》前3章
  - 劳虎&胭脂虎 《无废话XML》第1,2,5,8章
- 电子书到课件所在的群文件下载

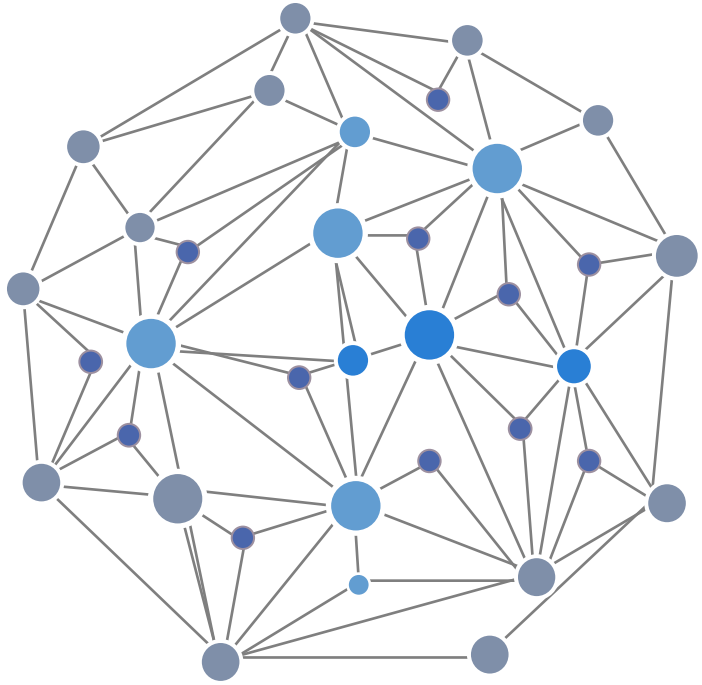
第三讲课后练习

- 编程作业 hw1
  - https://chenyv.gitee.io/webprogramming/homeworks/xml/index.html
  - 提交课程中心 met2.fzu.edu.cn

- 了解字符集和编码的基础知识
  - https://www.cnblogs.com/skynet/archive/2011/05/03/2035105.html
  - 《无废话XML》第三章

THANKS

本章结束



福州大学 计算机与大数据学院 软件工程系