



LANZHOU UNIVERSITY OF TECHNOLOGY

结课作业

题 目 XML技术描述

学生姓名 安阳

学 号 1817700216

专业班级 软件二班

学 院 软件学院

1.1 标记语言概述

人类总是在不断地改进对信息的表示与处理方式，科学技术发展到今天，计算机等各种信息处理调理已经将各种信息数字化，并逐步取代了手工的文档处理系统。

目前的各种电子文档不仅记录了所包含的数据内容，而且还记录了大量的手于排版、处理或印刷的格式信息。传统的文档格式化工作是由排版工人完成的，他们会给需要特殊格式的文字或数据作上标记。与此类似，当代的许多电子文档处理系统也是基于这种标记而工作的。在标记电子文档方面，最早成为国际标准的标记语言是结构化通用标记语言(SGML)，目前最流行的标记语言是超文本标记语言（HTML），而本书所介绍的则是有着广泛应用前景的可扩展标记语言（XML）。

1. 标准通用标记语言（SGML）

SGML(Standard Generalized Markup Language, 标准通用标记语言)，是一种定义电子文档结构和描述其内容的国际标准语言，是所有电子文档标记语言的起源，早在Web发明之前SGML就已存在。

IBM公司的有关技术人员在1969年就已开发出一种通用标记语言GML，用来解决不同计算机系统中文档格式不相同的问题。GML是一种可以创建各种数据标记的语言，能够方便地标识出任何数据集合的结构，因而成为IBM许多电子文档系统的基础。1986年 ISO（国际标准化组织）批准GML为国际标准，并正式定名为SGML。

SGML是用于定义其他语言的元语言，它以逻辑化和结构化的方式来描述文本化的文档，主要用于电子文档的创建、存储及分发。一个SGML文档由语法定义、文档类型定义和文档实例3个部分组成：语法定义部分为文档类型和文档实例定义了语法结构；文档类型定义部分定义了文档实例的逻辑结构和元素类型；文档实例部分则包含了该文档的所有数据元素，是SGML文档的主体部分。

正如其名称所言，SGML是一种采用标记来描述文档数据的通用语言，SGML定义了基本的语法，但也允许用户建立自己的元素标记。要利用SGML来描述特定文档，用户必须创造出一组适当的元素标记。例如，要描述有关书籍信息的文档，就可能需要命名并使用名为BOOK、PART、CHAPTER、INTRODUCTION、SECTION等的元素标记。

SGML是与平台无关的、结构化的、可扩展的标记语言。SGML文档包含了一系列DTD（Document Type Definition, 文档类型定义）内容，而在DTD中则定义了各种标记的含义，因而SGML的语法是可以扩展的。

SGML文档虽然已经在美国军方及航空业界使用多年，然而，W3C中的成员却认为SGML过于复杂和麻烦，难以学习和掌握，致使无法实现有效率地在网络上传递信息。SGML的过多弹性与其他一些特性，让编写负责在浏览器中处理与显示SGML信息的软件变得非常困难。鉴于这些原因，Web发明者欧洲粒子物理实验室的研究人员根据当时的情况，在SGML的基础上提出了超文本标记语言HTML。

1. 超文本标记语言（HTML）

HTML(Hyper Text Markup Language, 超文本标记语言)，是目前网络上应用最为广泛的语言，也是构成网页文档的主要语言。

1989年，欧洲粒子物理实验室的研究人员Tim Berners Lee 开发了一个基于SGML 的超文本简化版本，被称为超文本标记语言（HTML）。HTML 是SGML 的一个子集，继承了SGML的许多重要特点。

HTML 只使用了SGML 中很小一部分标记，并且为了便于在计算机上实现，HTML 可使用的标记是固定的，也不需要包含DTD，因而其语法是不可扩展的。HTML这种固定的语法使其易学易用，开发用来处理HTML文档的浏览器了也十分容易。正是由于HTML 的简单性和易用性，使Web技术ey6其应用得到了空前的发展，并很快从计算机界走向全社会，并最终走向千家万户。

HTML 使用各种定义好的标记来分别标识和格式化不同的网页元素，使得网页制作与Web信息发布变得非常简单。现在，一个普通人员无需了解HTML的语法就可以使用各种所见即所得的HTML编辑器，如Dreamweaver 、FrontPage 等，进行Web页面的创作。HTML已经成为最主要的Web页面语言。

然而，以目前的发展来看，HTML却存在着相当大的局限性。由于标准的HTML标记已经由W3C预先确定，不能根据需要自行定义，所以当描述具有各种复杂内容的文档时，HTML就显得力不从心。另外，HTML面向的是数据格式的描述，而非面向数据对象本身，因此HTML标记并没有给出数据内容的含义。

使用目前的HTML，网页文档开发者必须要对文档进行许多的调整才能兼容各种流行的浏览器。除此之外，由于浏览器不去检查网页中错误的HTML代码，因而导致Internet上有大量的文档包含了错误的HTML语法。这个问题越来越重要，于是W3C开始寻找Internet上有大量的文档包含了错误的HTML语法。这个问题越来越严重，于是W3C开始寻找解决方案，XML则是最好的选择。

1. 可扩展标记语言（XML）

XML(eXtensible Markup Language)即可扩展标记语言，它与HTML一样，都是出自于SGML标准通用标记语言。XML是Internet 环境中跨平台的、依赖于内容的技术，是当前处理结构化文档信息的有力工具。在W3C组织领导下的工作小组开发并支持XML技术，使用它来标记和处理通过互联网传输的文档信息。

XML具有如下一些特点：

XML是SGML的简化版本中，它继承了SGML的许多优点。

XML揭示了数据本身的含义，因而XML文档可用来清楚地描述、储存和共享各种数据。

XML是可扩展的，人们可以创建自己定义的文档元素以满足Web创作的需要。

XML是跨平台的，它不依赖于任何一种程序设计语言、操作系统和软件开发商。因而可以作为各种语言和各种系统平台数据交换的桥梁。

XML是结构化的，因此它不仅能够描述文档中各种数据的含义而且能够描述整个文档的逻辑结构。

XML是基于文本的，它可以用任何一种文本编辑工具来阅读和编辑修改。又因为它是基于Unicode的，所以是多语种兼容的。

XML的数据与其显示格式是分离的，因而同一个XML文档可以适用于各种不同的显示设备和显示方式。

因为XML文档具有良好的格式与逻辑结构，XML应用软件可采用简单的原则来区分文档中的数据与XML标记，因此程序员可以很容易地编写相应软件来访问XML文档。

1.2 HTML及其局限

要学好XML，首先应学习与了解HTML。Microsoft公司总裁比尔.盖茨在多年前曾经说过：“HTML 是我们的数据类型”。这句话预示了HTML 的发展前景。今天，互联网络和它最初用以表达信息的语言HTML已经获得了极大的成功，无论是公司、政府机构，还是个人网页设计者都已从中获益。由此看来，HTML 完全可以解决网络信息表示和交换的绝大部分问题了。但是，一些独具慧眼的人却已经认识到有必要重新发展一种新的标记语言。

1.2.1 HTML文档范例

HTML提供了一组固定的、预先定义好的元素标记，用来标注一般用途的网页元素，常用的元素标记包括：标题、段落、列表、表格、图片、超链接等。HTML 在创建普通网页时效果良好，大多数网页的代码都是由HTML标记的内容构成的。人们在上网浏览某个网页时，实际上是将该网页所对应的HTML文档下载到本地计算机中，再由浏览器软件对其进行解释和显示。

1.2.2 HTML 文档基本架构

HTML是一种颇为结构化的网页内容标记语言，使用各种不同的HTML标记符号来分别标识和设定不同的网页元素，这样的网页元素被称为HTML 元素。每个元素通常由开始标记（start tag）、结束标记（end tag），以及夹在这两个标记中的内容所组成。元素的开始标记和结束标记的名称相同（元素标记的名称也就是此种元素的名称），并且都需要用一对尖括号“<>”括起来，只是在结束标记的前面多了一个斜杠符号“/”。

许多HTML元素的标记中还允许加入若干个相关的属性，用以进一步对该元素的特征进行说明或限定。通常用“属性名称=属性值”的方式来设定某个元素的属性，属性值可以用半角的单引号或双引号括起来，也可以省略。例如要将图像元素<IMG>的来源属性SGC设定为Logo.gif文档，就可以使用<IMG SRC=”Logo.gif”>或者<IMG SRC=Logo.gif>的形式。在一个元素的开始标记中，对各个属性设置的前后顺序是可以任意的。

综上所述，一个HTML元素的形式可以表示为：

<标记名称 属性名称=“属性值”……>内容</标记名称>

一个完整的HTML文档是由文档头部和文档正文两个部分组成，其基本结构如下面的简单例子所示。

<HTML>

<HEAD>

<TITLE>

网页标题

</TITLE>

</HEAD>

<BODY>

</HTML>

说明：

<HTML>元素是HTML文档的最高层元素，原则上用它来指出HTML文档的开始结束。在<HTML>元素内部通常有<HEAD>和<BODY>两大元素，其他的元素都被嵌套在两个元素之中。

<HEAD>元素用来标识网页的头部，通常是向浏览器提供一些识别或补充的信息的。理论上在<HEAD>元素中至少必须包括<TITLE>元素，但在实际上并没有这么严格的限定。

<TITLE>元素用来指定浏览器所要显示的标题文字，原则上不应该省略，因为它还有其他用途。

<BODY>元素用来标识HTML文档的主体部分，它中间的内容即需要显示在浏览器窗口中的各种网页元素。

1.2.3 HTML文档常用标记

浏览器能够识别出HTML网页中每一个合法的HTML元素，并且知道应该如何在浏览器窗口中安排与显示它们。例如，浏览器通常会以最大字型来显示标记为<H1>的一级标题文字，以较小的字型来显示标记为<H2>的二级标题文字，以及用更小的字型来显示标记为<P>的文字段落内容。又如，浏览器会将包含在无序列表标记<UL>中的每一个<LI>元素加以顺序编号，并以缩进的段落来显示其内容。再如，浏览器会把超链接<A>元素的内容转换成含有下划线的热链接，让浏览者可以进行选择并加以单击以单击以打开所指向的网页。

1.2.4 HTML的局限

HTML页面的数据搜索耗时过长，HTML的链接机制不够完善，导致网页中的超链接目标在丢失后不能自动纠正。HTML无法表示非传统组件所构成的文档数据。HTML只能使用预先定义的标记，其可以使用的标记集合是固定的，可扩展性差。HTML本身的语法不够规范严密，严重影响网络信息传递和共享。HTML无法真正实现数据的交互。HTML不可重用。HTML要求文档过于完整

1.3 XML解决方案

XML代表了由W3C的XML工作小组所定义的可扩展标记语言。，该小组将XML描述为：“XML就SGML的一个子集……，其目标是让一般的SGML能够在网站上被请求、接收与处理。如同XML一样，XML应被设计成容易实际操作，并且能够与SGML及HTML共同合作的标记语言。”这就是说，XML是被特别设计用来在互联网上传递信息的标记语言，就像HTML一样，XML正在成为用来构建网页的标准语言。

1.3.1 XML的实现机制

XML的基本实现样式信息从文档中分离出来，然后将其放入相关的样式表（Style Sheet）文件中。譬如要标明一份在网上传输的订单，XML不是像HTML那样，对需要说明的数据采用在网页上显示的行、列位置及字体、字号等显示格式进行标注，而是根据这些数据所具有的含义，比如按订单号、商品名称、价格、尺寸、数量和颜色等性质来定义和标注这些信息。这样，浏览器程序就能够很容易地识别这个文件时客户的订单网页，并且能够按不同的样式表或者按不同用户终端的显示方式分别对其进行处理和显示。

一些经过精心设计的规则，使得XML能够作为互联网上信息表示、传送与交换的通用标记语言。首先，XML的标记总是成对出现，其功能就像我们通常使用括号一样，同时它们有具有引号的功能，可采用计算机中常见的树形结构进行多层次的嵌套，这就为程序员的编程处理提供了极大地方便。

总而言之，XML实质上是一种电子文档信息的自定义语言——任何人、任何行业都可遵循这些规则，自己定义各种标记来描述信息中的所有数据元素，然后通过一种被称为解析器的小型程序进行处理。信息组织与描述方式的这种看似简单的变化，其意义不仅在于改变HTML僵化的固定标记集，并且从根本上改变了传统网页的信息构造方式，创造了一种让所有计算机系统都能够进行“智能”交流的语言，使得基于信息传输的互联网在信息获取、检索和交换方面发生一系列重大变革，为整个互联网的智能化奠定了基础。

1.3.2 XML文档实例

在IE浏览器中显示的XML文档几乎和源文档的格式与内容一样，这是因为对于XML文档本身而言，通常人们只关注它的数据内容，而它的显示形式是 交给CSS或XSL来定义的。如果XML文档包含了所链接的某个样式表，IE浏览器将只显示XML文档中的数据，同时根据这个样式表中指明的显示规则来安排这些数据的显示格式。

1.3.3 XML的官方目标

（1）XML应该在Internet上直接可用（2）XML应该支持各类不同的应用程序（3）XML应该与SGML兼容（4）应该很容易编写处理XML文档的程序（5）XML中可选功能的数目应该维持在最小值，理想情况是零（6）XML文档应该是易读的，并且是清楚合理的（7）XML的设计应该能在一定时间内迅速地完成其阶段性的任务（8）XML的 设计应该合乎规范并且简洁（9）XML文档应该容易建立（10）将XML标记简化是最不重要的。

1.4 XML的特点

1.4.1 XML的优势

XML具有良好的可扩展性，方便的可扩展性，数据内容与其形式的分离，便于信息的检索，可以轻松地跨平台应用，支持不同文字不同语种间的信息交互，为科学实践与交流提供捷径，大大改善网页的超链接性能，便于信息的长期保存，适合面向对象的程序开发

1.4.2 XML 与Java

有了XML和Java技术的合作，许多人物可以在客户端完成，更多的应用软件处理起来也不会占用太多的网络通信量，从而使得网络更加畅通快捷，同事客户端也可以运行多个应用软件。

1.4.3 XML的不足

XML的数据是以树形结构存储的，但数据的插入与修改却比较困难；XML文档的内容是完全以文本形式表现与存储的，XML文档中标记与数据的字符化表示与采用二进制方式表示相比；XML文档作为数据集合使用时，相当于一个数据库，却不具备数据库管理系统那样完备的管理功能和查询功能；XML不能替代HTML；由于XML是一种元标记语言，任何组织、任何人都可以利用它来定义新的标记和标准，从而将产生一定的混乱与冲突。

1.5 XML的应用

1.5.1 XML的应用领域

XML在语法结构上的优越性，使得它可以广泛地应用于信息表现、信息交换、信息存储、Web服务、电子商务、网络出版等多个领域。随着现代信息网络技术的发展，XML必将在更多的领域中发挥重要的作用。

数据交换；Web服务；Web集成；电子商务；数据库应用；网络出版；特定文档的组织；系统配置文件。

1.6 XML 相关技术

文档类型定义（DTD）；层叠样式表（CSS）；可扩展样式表语言(XSL)；数据源对象（DSO）；文档对象模型（DOM）；XML Schema；Xlink 与Xpointer。

小结：

XML是由W3C推出的新一代网络数据传递和交换的标准，是Internet环境中跨平台的、依赖于内容的技术。W3C的全称为World Wide Web Consortium，即万维网联盟，其网址为<http://www.w3.org/>。

当代的许多电子文档处理系统是基于某种标记而工作的。在标记电子文档方面，最早成为国际标准的是标准通用标记语言SGML，目前最流行的标记语言是超文本标记语言HTML，而有着广泛应用前景的则是可扩展标记语言XML。

HTML提供了一组固定的、预先定义好的标记用来标注网页元素。常用的元素标记包括：标题、段落、列表、表格、图片与超链接等。HTML在创建普通网页时效果良好，然而目前的发展来看，HTML仍存在着相当大的局限性。

XML是一套完整的解决方案，与之相关的一系列技术包括：用于规范和验证XML文档的DTD与XML Schema技术，用于格式化显示XML文档的CSS 和XSL样式表技术，用于把XML文档作为数据对象进行显示和编程处理的DSO和DOM技术，以及在XML文档中使用XLink、XPath和XPointer链接技术等。