《情报科学》2001 年第 9 期第 964-971 页，续 983 页

基于 XML 的信息组织与处理：2. 应用技术

张晓林

文摘：本文介绍基于 XML 的基础数据与应用文献标记、元数据及知识体系标记、基于 XML 的应用领域信息处理与交换框架建设，并分析 XML 在图书情报领域中的应用以及 XML 对图书情报领域的挑战。

关键词：XML，元数据，信息组织，信息处理，信息处理与交换框架

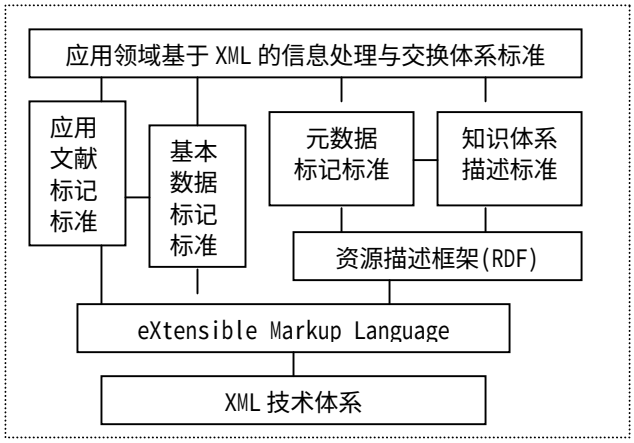
XML-based Information Organization and Processing: 2. Applied Technologies

Abstract: The paper describes XML-based applied techniques, explores the XML-based information processing and exchange frameworks in application fields, analyses the applications of XML in and its challenges to the library and information service field.

Keywords: XML, Information organization, Information processing, Information processing and exchange framework

1. XML 技术体系和基于 XML 的应用技术

正如我们在文献[2]中指出，以 XML 为代表和基础的 XML 技术体系[1-2]正日益成为数字化网络化信息环境中 对信息进行组织、处理和交换的基础。通过 XML 技术体系，我们能够定义和标记由任何数据类型组成的信息集合模式(文献模式)，定义和标记复杂形式的元数据和知识组织体系，定义和标记这些信息集合模式及其片段的命名、抽取、链接、合并、集成，定义和标记在不同信息集合模式间进行转换和在输出介质上呈现这些信息模式的机制，定义和标记这些信息集合的数字签名机制和加密机制，定义和标记各种信息系统对基于上述信息模式的信息资源的开放式处理界面，定义和标记对基于上述信息模式的信息资源进行查询的机制。



由于 XML 技术的开放性、灵活性、机器可处理性和可扩展性，它们正被越来越多应用领域接受为信息组织和处理的基础工具，并由此出现一系列应用技术及基于 XML 技术的应用领域信息处理和交换基础框架[3](图 1)。

本文选择介绍这些方面的有关技术、标准和系统，着重分析其作用和意义，请欲了解细节的读者阅读参考文献中给出的具体标准。

2. 基于 XML 的基础数据和应用文献标记

XML 是基于文本形式的标记语言。为了表达丰富的信息内容，需要定义用 XML 来标记其它的复杂数据类型的标准方式，为此 W3 联盟(WWW Consortium)和有关应用领域提出了一系列标准。

2.1 HTML 的 XML 化语言 XHTML[4]

XHTML 是将 HTML 文件转换为规则化 XML 文献的标准方法。实质上，XHTML 将 HTML 文件定义为 XML 文献，建立相应文献类型定义(DTD)，并建立相应转换规则(例如将缺少结束符、不规则嵌套等不符合 XML 规则的标记转换为规则方式)，从而将其转换为保留基本 HTML 标记的规则化 XML 文件，可由 HTML 或 XML 浏览器释读。经过转换的文件具有 MIME 类型 text/html 或 application/xhtml+xml。

每个 XHTML 文件包含 XML 声明、XHTML DTD 链接和 HTML 文件三部分。XHTML 定义了三种 DTD，一是严格 DTD (XHTML Strict)，严格按 XML 要求定义允许的 HTML 元素和属性，并将所有涉及屏幕表现格式的元素(字形、 颜色、字体等)用 CCS 层叠格式单语言表示，形成规则 XML 文献；一是过渡 DTD(XHTML Transitional)，在按 照 XML 定义 HTML 元素时允许一定屏幕表现格式标记，以便向后兼容；一是框架 DTD(XHTML

Frameset)，用 XML 对 HTML 框架元素进行定义和转换。XHTML 可通过模块化或子集方式将 XHTML 分为一系列小的元素集(并可重组这些模块或子集)，用于不同设备或应用平台(例如手持设备)；还可根据应用需要利用这些模块、子集及 DTD 定义来建立新的文献协议(Document Profiles)， 从而用标准 HTML 元素及其它元素构建新文献结构。

2.2 可伸缩矢量图像标记语言 SVG[5]

SVG(Scalable Vector Graphics)定义用 XML 语言表述二维图像的标准方式，该图像可容纳矢量图形、 点阵图像和文本。SVG 用 XML 语言定义基本元素 svg(svg element)，并定义包含的基本矢量图形元素(直线、 长方形、圆、圆弧、椭圆、圆弧、多边形、路径及闭合路径等)、文字元素及其样式和路径描述、图像对 象元素(可由多个图形组成，或由若干图形、点阵图像、文本等构成，有名称、样式属性，有并行的标题和描述元素，可被 SVG 文件其它部分链接和调用来产生新的图形对象)、图像坐标体系及其转换和旋转机制、图形填色滤色及色彩梯度表示方法。SVG 还定义对图形图像在位置、比例、色彩、大小、移动路径等方面按一定时序进行动画处理的机制(包括有关操作事件及相应过程控制)，定义为图形图像建立链接的方式，定义各种元素的应用程序界面(DOM 界面)和嵌入脚本语言对有关元素进行处理的方法，从而使 SVG 图像可以动态和交互产生和处理。

SVG 元素可作为单独 XML 文献(MIME 类型为 image/svg+xml)，也可作为元素嵌入 XML 文献中。SVG 元素还可将其它 svg 元素表示的图像链接称为自己组成部分。SVG 可允许 SMIL(同步多媒体合成语言)将 SVG 内容作为多媒体合成文献的一部分，也将与 SMIL 模块和其它 XML 工具共同构建动画效果。

作为基于 Web 的图像表达和传递机制，SVG 可伸缩性表现在既可对图形图像进行多视角多层次表现和处理，又可链接其它内容对象和内容描述语言、其它工具和应用界面来进行灵活处理和扩展。

2.3 同步多媒体合成语言 SMIL[6]

SMIL (Synchronized Multimedia Integration Language)定义用 XML 语言描述同步多媒体合成的标准 方式。利用 SMIL，人们可将一组独立的多媒体对象(包括动画、声音、图像、文字、视频图像等)合成为同步多媒体演示，准确描述演示的时序行为和空间布局布局。由 SMIL 描述的演示文件称为 SMIL 文件，MIME 类型建议为 application/smil+xml。SMIL 文件还可嵌入其它 XML 文献中。

SMIL 由 HEAD 和 BODY 两部分组成，在 HEAD 部分主要定义空间布局及元数据和选择开关，空间布局包括根演示窗口和多个媒体演示窗口，SMIL 可定义它们的大小、位置、背景颜色及媒体演示窗口的叠放顺序等。 SMIL 的 BODY 部分包含了演示内容和同步机制，SMIL 定义了两个同步元素 par 和 seq，其中 par 元素将多个媒体对象并行演示，seq 元素将多个媒体对象顺序演示，当然两者都可对其中某些媒体对象规定具体的演示开始和结束时间及演示长度。SMIL 在同步元素中通过超链元素来链接要播放的各个媒体对象。

SMIL 还定义了选择开关 switch 元素，可根据系统或媒体对象的有关条件在演示时选择演示其中的某一组媒体对象，例如根据媒体对象语言类型演示不同语言的文字，或根据线路带宽演示不同分辨率的图像。

2.4 数学标记语言 MathML[7]

MathML(Mathematical Markup Language)是基于 XML 语言的描述数学公式结构和内容的标准方法，支持 基于 Web 的对数学信息进行表达、传递和处理。我们知道，任何数学表达式都可逐层分解为由一定运算符组合的子表达式，直到最基本的数学元素或符号。因此，MathML 设计了相应的元素和标记方法来表示这些有限个数的运算符和数学元素，并通过它们的组合来表达任意数学表达式。

MathML 通过两类元素来标记和表示数学表达式，一是表征元素(Presentation elements)，一是内容元 素(Content elements)。在使用表征元素时，MathML 将数学表达式作为由数字、字母和数学符号等基本符 号组成的可视二维结构，定义有哪些基本符号、如何标记、怎样组合来形成数学表达式，如上标、下标、分号、括号、根号、矩阵等。一个复杂公式可标记为由一定基本符号连接的若干子表达式的组合。MathML 定义了 30 余个表征元素，包括基本符号元素(Token Elements)和符号组合元素(Layout Elements)。在使用 内容元素时，MathML 将数学表达式看成是由抽象数学对象构成的集合，每一个表达式不再是简单的可视符号组合，而是具有实际语义的内容结构。因此，它按照数学含义定义了 120 个内容元素，例如乘、乘方、微分、正弦等，覆盖代数、几何、微积分、线性代数、

统计学、逻辑学、矢量、集合等领域。通过这些内容元素及相应的结构组合元素(具体规定内容元素组合方式)，MathML 可明确标记和描述数学表达式的内容含义和组成结构，而不仅是对数学表达式作可视化表征。在实际中，表征元素和内容元素可在一定条件下混同使用。MathML 还定义了链接外部数学符号标记、定义新的数学标记方式等的机制。

按照 MathML 标记的内容可嵌入 XML 和 XHTML 文件中。可解析 MathML 标记的浏览器目前已存在。

除了上述几种外，目前还有其它基于 XML 的基础数据标记语言，例如声音合成标记语言 SSML[8] (Speech Synthesis Markup Language)、音乐标记语言(MusicXML)[9]、地理标记语言(Geography Markup Language， GML)[10]、化学标记语言(CML-XML)[11]等。

2.5 应用文献模式标记

由于 XML 的灵活性和可扩展性，许多领域用开始 XML DTD/Schema 来定义本领域有关文献的标准结构和标 记方式，例如金融信息领域就提出 Trading Partner Agreement Markup Language (tpaML)、Extensible Financial Reporting Markup Language (XFRML)、Extensible Business Reporting Language (XBRL)、 Financial Products Markup Language (FpML)、Market Data Markup Language (MDML)、MarketsML、 swiftML for Business Messages 等(参见[3]给出的 XML 门户网站)。许多领域开始对多种具有相似功能的文 献标记模式进行整合，最后形成的模式将成为该应用领域的标准模式，在开放式 XML DTD/Schema 登记系统 登记，可在具体 XML 文献中作为文献模式或名称域引用。

3. 基于 XML 的元数据与知识结构标记

所谓元数据，包括对信息实体及信息集合的各方面特征和管理使用要求等进行描述的数据，可简单分为三个层次：一是对具体信息实体(例如文献)进行描述的元数据，包括内容著录数据、技术指标数据、日常管理数据、使用控制数据、知识产权与内容评鉴数据等；二是对信息集合进行描述的数据，涉及网站、信息频道、数字图书馆、数据库、档案库等实际或虚拟的信息集合，其元数据包括内容体系数据、使用管理数据、知识产权管理数据、隐私保护管理数据、内容评鉴数据、保存管理数据等；三是对信息内容及信息实体或信息集合关系进行描述的数据，主要包括词表、语义网络和应用知识体(Ontologies)数据。这里将前两类数据称为元数据，将后一类数据称为知识结构数据。

3.1 资源描述框架(RDF)

目前，许多应用领域开始建立自己的元数据标准[12]，然而为了在网络环境下自动识读和交换这些源于不同目的和历史、应用于不同领域、具有不同语义的元数据，需要一个统一的描述框架和标记语言，W3C 的资源描述框架(Resource Description Framework，RDF)[13]就是这样的“宏”标准。RDF 认为，一个具体元 数据实际是关于特定资源的特定属性的取值声明，是一个由资源、属性、属性值构成的三元关系模式，例如“网页 ABC.com/XYZ/abc.html(资源)的制作者(属性)是 John Smith(属性取值)”。不同应用领域的元数 据可能定义不同的属性集合以及这些属性的取值范围，但它们都只是资源、属性、属性值三元关系模式的具体体现而已。鉴于此，RDF 定义了用 XML 语言来描述这种三元关系的基本方式，从而建立所有元数据定义和交换的基础平台。例如，

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
        xmlns:s="http://description.org/schema/">
   <rdf:Description about="http://www.w3.org/Home/PubText/">
    <s:Creator>Ora Lassila</s:Creator>
   </rdf:Description>
</rdf:RDF>
```

其中，rdf:Description about 通过 URI 指出所描述的资源，s:Creator 标记所描述的资源属性并可用名 称域链接来定义属性名，而属性标记符之间的值就是属性取值。通过这种基本结构，RDF 可用 XML 语言标记任何元数据，基于 RDF 和 XML 的浏览器可解析相应的元数据。对于更复杂情况，例如，属性值本身也是资源 (例如该网站作者 Lassila 本身又可由邮件地址和主页代表)，属性本身可能还有自己的限定属性(例如来自什么主题词表或取什么重量单位)，元数据描述语句本身又可能有自己的限定(例如是谁做出的这个描述、这个描述的可靠性等)，这些关系都可用 RDF 三元语句来进一步描述。RDF 还定义了集合型元数据描述语句 (Containers)，用以描述可取或可选多个属性值的资源属性，例如无序表

(Bag)、有序表(Sequence)、选择表(Alternative)。

需要指出，RDF 本身并不直接定义具体元数据，而是定义元数据与资源最基本关系的基础描述模式。具体元数据名称和结构往往由实际应用领域定义，RDF 通过 XML 名称域引用其中任何合适的元数据元素作为属性名称来描述相应资源。这种元数据标记方式独立于任何具体的元数据格式，可以用标准方式标记和交换任何具体元数据，又可引用和集成多个元数据格式(例如内容描述、内容评鉴、接入控制、知识产权保护、隐私保护、数字签名、支付控制等) 来灵活和全面地描述元数据及其管理使用控制要求。同时又由于三元 关系模式的简单性和 XML 语言的通用性，可在任何基于 XML 平台上方便地解析用如此标记的元数据，从而提供了统一和机器可读的元数据标记和交换机制。一些元数据项目已经开始试用 RDF 标记自己的元数据，例如 Dublin Core[14]和 PICS[15]。用 RDF 描述的元数据可嵌入 XML 或 HTML 资源文件，可作为外部 RDF 文件单独存在于 元数据库，可作为外部 RDF 文件被资源文件用 HTML/LINK 元素链接，也可将资源文件封装在 RDF 文件中。

3.2 RDF 模式语言(RDF Schema)[16]

RDF 可通过 XML 名称域方式将元数据元素名称(资源属性)与对应定义文件链接起来，从而可解释这些元数据元素的基本定义。但是，元数据元素及其子元素间可能具有复杂的多层的类属关系或其他形式的语义关系，这些元素本身往往拥有一定属性，这些属性之间可能又有复杂的类属关系，而且这些元素或元素属性可能限定应用于特定类别的资源和特定的取值范围。描述和理解这些关系，对于计算机对元数据及它们所描述的资源的自动理解和智能处理至关重要。为此，W3C 通过 RDF 模式语言(RDF Schema，简称 RDFS 语言)定义了用 RDF/XML 来描述元数据模式的标准方法和词汇。

从 RDFS 角度，任何元数据可看成是一个描述特定资源实体(如图书、网页、汽车、系统)及其属性的概念。这些概念本身往往组成一个层级类别体系，即具体元数据值只是某个概念类别的实例，而该概念类别可能是某上层概念类别的子类，而这个上层类别又可能是更上层类别的子类。例如狗是哺乳动物的子类，而哺乳动物又是动物的子类。这些概念类别所代表的实体具有一定的属性，这些属性本身间又可能有一定的层级关系。RDFS 通过 rdf:type 定义元数据概念隶属的概念类别(资源实体或属性，或者某个概念类别)，通过 rdfs:subClassOf 和 rdfs:subPropertyOf 定义其与父概念对象的关系，通过 rdfs:range 和 rdfs:domain 定义概念对象所允许的取值范围和应用类别。例如

```
<rdf:Description ID="大学生">
  <rdf:type resource="http://www.w3.org/2000/01/rdf-schema#Class"/>
  <rdfs:subClassOf rdf:resource="#学生"/>
</rdf:Description>
```

利用 RDFS 语言，元数据设计者可定义所描述的资源类别和属性类别及其词汇，可定义这些对象或属性类别的类属关系及对象与属性间相互关系，可进一步定义这些资源对象、属性及属性应用类别范围和取值条件，从而以计算机可理解的标准方式描述元数据(如 MARC、EAD、Qualified Dublin Core)的语义内容和 元素关系结构。除了一般元数据外，RDFS 语言还具备必要语义工具和能力来定义网站资源图(sitemaps)、专业词汇表、叙词表、分类表等逻辑知识体系。由 RDFS 语言定义的元数据体系称为 RDF 元数据模式，利用它们来描述具体资源的元数据的文件是对应 RDF 元数据模式的实例，称为 RDF 元数据文件。与 RDF 类似，RDFS 语言并不定义任何具体元数据模式，而是定义描述这些元数据模式的标准方式。所形成的 RDF 元数据模式本身是 RDF 文件(即是规则的 XML 文件)。在任何可解析 XML 的平台上，应用系统即使事先不知道对应元数据模式，在释读 RDF 元数据文件时，可调用被链接的 RDF 元数据模式来理解元数据元素的含义及其相互语义关系，从而利用它们(尤其是其语义关系)进行处理和推理。而且，人们可同时链接和利用多个分布的 RDF 元数据模式来多角度多层面地描述一个资源，可以共享和重用这些 RDF 模式，甚至可利用若干 RDF 模式来方便地定义新的 RDF 模式，从而使元数据的定义和利用更具灵活性和可扩展性。

3.3 XML 主题图(XML Topic Maps，XTM)[17]

主题图(Topic Maps)有两个含义，一是特定主题概念关系体系(例如叙词表)，一是一定资源集合主题 内容的结构化表现(例如百科全书主题索引或网站 Sitemaps)。主题图独立于应用技术平台，可描述所涉及的主题词汇、这些主题间的关系以及这些主题与具体资源的联系，可“标引”信息资源并建立相应索引或交叉参照，可链接复杂主题范围的分布式资源

来建立虚拟知识体系，可通过主题概念与资源的不同链接在同一资源集合基础上建立面向不同主题或不同用户的资源界面。XML Topic Maps(XTM)就是基于 ISO 13250 标准，定义用 XML 语言描述和标记主题图的标准方式。由 XTM 标记的主题图是 XML 文件，称为 XTM 主题图。

XTM 用主题(topic)代表具体的实体或概念对象，这些主题可被一定信息资源描述、讨论或提及。XTM 规定这些主题在主题图中具有唯一的确认名(ID)、具有一个基准名称(baseName)、可以是另一个主题的实例或子类(instanceOf)、可以出现(occurences)在若干个不同形式的用 URL 表示的信息资源里。XTM 定义相应的元素及用这些元素来表示主题的具体语法。此基础上 XTM 定义描述主题关系的关联元素 association，一个关联元素可能包含若干个主题(members)，这些主题按照特定角色(roles)发生特定的相互关系，例如莎士比亚(主题) 作为作者(角色)与名为哈姆雷特(主题) 的戏剧(角色)之间发生“写作”(written_by)关系。这些关系类别可包括隶属关系、实例关系、逻辑关系等，本身可作为主题在主题图中定义。因此，XTM 主题图就是用 XTM 标记的一组主题及其相互关系和这些主题所链接资源的集合。一个 XTM 主题图可被用来以不同形式描述和链接不同资源集合。反之，同一资源集合也可被不同 XTM 主题图以不同形式描述和链接。

就象不同主题词表可能为同一实体定义了不同主题词一样，不同 XTM 主题图可能为同样的实体或概念在不同应用环境下定义不同的主题。为明确主题的含义，XTM 规定可用其它主题、外部名称域(namespaces)、外部公开发表的主题定义(subjectIndicator)来定义某个主题的应用范围(Scope)。进一步地，XTM 可利用主题的基准名称和范围限定来比较和合并相同主题、甚至相重合的主题图。XTM Processing Requirements 定义了比较和合并的具体条件和操作过程。

3.4 知识体系标记与 Semantic Web

利用 XML DTD 或 XML Schema 可以解析 XML 文献的内容元素，利用 RDF 可以解析元数据元素，并据此对 XML 文献或元数据文件进行检索、过滤、转换等处理。但是，这些标记元素的含义取决于具体的应用领域，例如 TITLE 在出版领域代表书刊题名，在行政领域则可能代表职衔；而且，TITLE 与 PERSON 在这两个领域也具有不同关系。明确地定义 XML 内容元素和元数据元素在特定领域的含义、明确定义这些元素在该领域的语义关系，并用计算机可解析的语言来标识和交换这些定义，将使计算机具备理解逻辑内容和语义关系并在此基础上进行智能推理的能力，这就是 Semantic Web 的基本目标[18]。为此，不仅要有 XML DTD/Schema 和 RDF 来规范标记文献内容元素和元数据元素，而且需要 RDFS 来定义和标记具有复杂关系结构的元数据模式。但即使 RDFS 语言也过于简单，人们正设计和试验专门用来描述应用知识体的标记语言。所谓应用知识体，指关于特定领域的概念体系及其相互关系的集合，一般包含概念类别的层级体系及类别组合关系(例如交集、并集、补集)，概念类别语义关系，概念属性及其层级关系，概念实例化关系及概念属性取值限制和传递转换规则等，以及关于概念对象及其关系的推理规则。应用知识体还可能包括应用领域的活动流程或具体应用的处理流程，例如电子商务体系或医疗处理流程。应用知识体标记语言就是为人们定义具体领域知识体提供标准的语言工具和标识语法。这方面典型的例子是基于 XML、RDF 和 RDFS 的 DARPA Agent Markup Language (DAML)[19]，应用知识体交换语言(Ontology Interchange Language)[20]，以及 Simple HTML Ontology Extension(SHOE)[21]，两者均可用于定义和描述应用知识体，形成用 XML 语言标记的应用知识体文件。人们可利用这个知识体文件来为 HTML 或 XML 文献加上符合该知识体的内容元素或元数据元素(Ontology-based Annotation)， 有关智能代理可在这个文件的支持下分析 HTML 或 XML 文献，理解其准确内容含义，理解不同 文献或同一文献不同部份的语义关系，从而实现自动理解和推理。

4. 基于 XML 的应用领域信息处理与交换框架体系

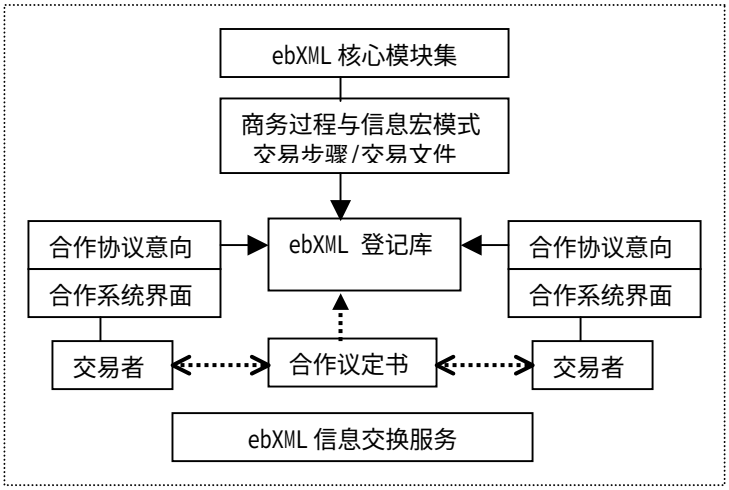
由于 XML 语言及其标准技术的开放性和可扩展性，许多领域正积极建立基于 XML 的信息处理与交换框架体系，促使整个应用领域或某一业务流程中所有各方都利用 XML 来定义、组织和交换信息。

4.1 基于 XML 的电子商务信息框架(ebXML)

ebXML(Electronic Business XML)[22]是联合国贸易发展与电子商务中心(UN/CEFACT)和促进结构性信息 标准化组织(OASIS)共同发起的一项国际性研究与发展计划，试图建立一

个基于 XML 技术体系的开放性全球电子商务信息交换框架。该计划有关工作组已提出基本技术体系和若干标准。

ebXML 认为，任何商务活动都体现为由若干个交易伙伴参与的商务过程(Business Process)，每个商务 过程涉及若干交易步骤(Transactions)，每个交易过程又交换若干商业文件(business documents)。ebXML 规定了图 2 所示的信息处理交换框架。其中，商务过程与信息宏模式(Business Process and Information Meta Models)定义对应商务过程所涉及的交易伙伴及其角色、相互关系和责任，定义这个过程将涉及的对 应交易步骤，定义这些交易步骤所需要的由不同角色交易伙伴提供的各种交易文件，并定义这些交易文件的具体内容和格式；这些模式及过程由各个行业或应用领域用统一模式化语言 UML 描述，用 XML 语言标记，有关交易文件用 XML 模式语言标记，所有标记文件各被赋予一个唯一标识号，存储于开放式的 ebXML 登记库(Registry)。edXML 将提供核心商业过程模块、核心交易文件模块和核心商业过程功能来帮助各应用领域定义宏模式。交易伙伴(Trading Partner)按照 ebXML 方式 定义自己的合作协议意向(Collaboration Protocol Profile)，具体说明自己愿意参与和支持什么商务过程及相应的交易步骤和交易文件，说明自己支持这些商务过程的系统界面细节，并提供自己的其它信息；这些意向文件用 XML 语言标记，被赋予唯一标识号，存放于 ebXML 登记库。交易伙伴间可根据双方的意向文件建立合作议定书(Collaboration Protocol Agreement)，具体规 定双方同意参与的商务过程、权力与义务、及采用的交易步骤和文件，这些议定书可存放于 ebXML 登记库。ebXML 登记库提供各种标记文件的公共存储和开放式查询服务，并可通过分布式登记库体系支持跨行业跨国家的开放式查询服务。各交易者系统对登记库的查询及交易者间的信息交换由 ebXML 信息交换服务(ebXML Messaging Service)支持，它在公共通讯协议(HTTP、FTP 等)基础上定义有关信息交换服务，并用基于 XML 的特定格式封装被传递的信息。



按照这样一个框架，有关的商务过程、交易步骤、交易者及其条件、交易关系及其限定条件、交易系统界面等都利用 XML 技术体系详细定义，任何交易者都可通过 edXML 登记库查询自己所需要的商务过程和交易者，然后与合适交易者建立合作议定书，并按照商务过程宏模式来建立所要求的系统界面，就可建立与交易伙伴间的电子商务系统。交易者还可制作自己的合作协议意向，提交 ebXML 登记库存储，供其它交易者查询和连接。

4.2 其它基于 XML 的信息处理与交换框架

英国电子政府信息处理框架(e-GOV/UK)[23]：英国政府提出的电子政府计划中明确规定，将 XML 作为所有 公众系统电子信息组织和表示的核心标准，将采用 XML 作为政府信息互操作和集成战略的基石。这项计划具体规定，用 XML 语言和 XML 模式语言来建立政府信息处理所涉及的所有文献格式，将 XML 和 RDF 作为相应元数据和知识体系的描述语言，将 XML、XSL 和 DOM 作为数据表示和转换的基础工具，从而保证各政府机构及密切相关机构能够按照统一标准和自动方式建立、识读和交换信息。该计划进一步规定，在政府信息范围内所有新系统必须采用上述标准，要接入有关政府信息网络或门户的旧系统也必须与上述要求兼容。

电子文件档案库(Electronic Record Archives)[24] ：美国国家档案管理局提出的电子记录档案库计划 也建议，采用 XML 文献类型定义或文献模式来定义电子记录格式，采用 XSL 来在输出介质上表现电子记录，采用 XML 主题图来反映档案集合内部结构和关系，采用 XSLT 来转换各种电子记录。

类似应用体系还有临床文件结构(Clinical Document Architecture，CDA)[25]，地理与空间数据协调体 系[26]，学校信息互操作框架(Schools Interoperability Framework)[27]，共享课

件对象参考体系(Sharable Courseware Object Reference Model)[28]、基于于 XML 的新闻处理与交换框架(NewsML)[29]等。我们相信，随 着 XML 技术体系的进一步完善，它将成为各个应用领域普遍的信息处理和交换基础。感兴趣的读者可参见引文[3]了解更多的细节。

5. XML 在图书情报领域的应用

图书情报系统是以信息组织、处理和传递为基础能力的服务体系，有效地采用 XML 技术体系将能显著地提高这些能力，也非常有助于与其它领域信息体系的交互和集成。目前，XML 技术在图书情报领域的应用已得到普遍的重视，并在许多发面进行了有益的探索，包括：

(1) 利用 XML 直接定义和标记各种文献格式，尤其是文本文献。传统地，许多图书情报系统采用 PDF 格 式或简单文本格式，但今后趋势将是采用基于 XML 的或可与 XML 转换的标记语言和格式。例如，英国图书情报网络办公室(UKOLN)规定文本数据必须采用 HTML、SHTML、XML 标记语言，矢量图像应该采用 SVG 语言，美国国会图书馆也规定采用 SGML 标记文本数据。广泛应用的文本数字编码格式(TEI)一开始就为小说、戏剧、诗歌、非小说著作等定义了对应的 SGML/DTD，现在又开始建立对 应的 XML 文献类型定义[30]。

(2) 利用 XML 标记各种交换格式。例如，开放数字资源库系统(Open Archives Initiative)定义了用来 描述 Dublin Core、MARC、RFC 1807 等元数据记录的 XML 模式。所有参加该系统的分布式数字资源库，无论 其内部采用什么格式标记和存储这些元数据，都必须用 XML 模式向检索服务器提交有关元数据，而检索服务器也以 XML 模式向用户界面提交元数据。这样，用户可通过 XML 方式来检索用任何方式实际存储和记载元数据的任何数字资源库[31]。另外，美国国家医学图书馆从 2001 年起采用 XML 格式来传递 Medline 数据[32]，欧洲可视档案项目规定采用 XML 格式在多个国家图像档案库间交换元数据[33]，还有许多出版商业在试验采用 XML 格式传递文摘索引数据。

(3) 利用 XML/RDF 来定义和描述各种元数据与知识体系模式。我们已提到，已可采用 RDF 来表示 Dublin Core 和 PICS，用 XML 来表示主题图。人们也积极试验用 XML/DTD 表示档案编码描述(EAD)[34]、 Making of America 二期项目所有数字化对象的元数据[35]、共享课件元数据[28]等。已有人提出，原来用专有格式、只有图书馆系统能识读的 MARC 应该被转换为 XML 格式，以便其它非图书情报机构能够利用 XML 技术要处理 MARC 记录[36]。由于 XML 的开放性、可扩展性和机器可处理性，它将成为元数据的主要描述和交换语言。

(4) 利用 XSLT 对各种文献或元数据进行转换。不同系统为了不同应用需要在不同时候定义了许多文献 格式和元数据格式，要求它们全部统一到一种格式上既不科学也不现实。现在，利用 XML 技术体系、XSL 格式单和 XSLT 转换语言，可以将任何格式的文献或元数据自动转换成所需要的其它格式。目前在图书情报界这方面试验主要集中在将 MARC 转换为 XML 格式，或利用 XML 在 MARC、Dublin Core、VRA 等之间进行转换[37]。

(5) 基于 XML 的数据挖掘。基于 XML 的文献或元数据都是一种结构化数据，可以利用 XML/DTD、XML Schema、XML 名称域及 XSL 等来自动识别文献结构和解析文献内容，挖掘有意义的结构信息或主题内容，从而支持对文献的自动识别、过滤、分类、标引等操作，也更深入地针对文献内容进行检索。例如斯坦幅大学图书馆正试验从 XML 格式的电子期刊文献中自动提取和编制著录记录[38]。实际上，XML 技术体系提供了对文献进行灵活解析和重组的有力方法，在此基础上信息服务系统可根据用户的要求来动态地获取、组织、抽取、转换、集成、传递信息[2]。

(6) Semantic Web 与网络资源智能检索。显然，前述 RDF、RDFS 和 Semantic Web 等技术将为基于概念的 智能检索和推理机制提供必要的技术条件，可帮助我们充分利用各应用领域内在知识结构来组织网络信息资源和提高检索效率，并为根据用户需求来过滤、转换、抽取、重组和传递信息打下坚实的知识化基础。

6. XML 技术对图书情报领域的挑战

根据文献[2]和本文的分析，我们相信 XML 技术体系为网络化数字化信息环境提供了新的信息组织与处理的核心能力，从而保证信息系统能够开放地自动地甚至智能地进行用户所要求的各种复杂信息组织和处理操作，允许以前所未有的灵活性和深度对信息进行动

态加工来提取和组织知识，并能在各个应用领域或不同信息系统间有效实现无缝交换、虚拟集成和互操作性。

掌握 XML 技术及其应用，不仅能促进我们有效地组织数字化资源和网络化信息服务系统，还将保障我们对飞速增长的基于 XML/HTML 的网络信息资源的处理能力和与其它领域信息系统的互操作性，帮助我们有效参与各应用领域基于 XML 的信息处理交换体系、充分发挥和扩展我们作为信息组织与处理专家的作用[36]。

但是，要做到这些我们也面临很多挑战。显然，XML 技术体系和基于 XML 的应用技术对我们来说是一个新的思维方式和知识领域，我们需要进行全面的再教育。但也许更为严重的是现有系统和思维的限制。一方面，在图书情报领域专有数据格式和数据处理机制占主流地位，即使标准 MARC 也缺乏严重开放性，各类信息(包括数字信息)常在具体系统条件限制下用专门语言定义组织为内部结构和格式，不同类型和不同层次的数据(例如基本文本数据、图像数据、元数据、词表数据等)常常在概念、技术、甚至物理上被定义和组织成不同格式和形态，难以有效进行机器支持的检索、解析、处理和交换，难以进行跨文献单元、数据类型、数据层次和系统范围的信息挖掘、抽取、综合分析描述、转换，也难以与其它领域(例如出版发行、地理信息处理、电子商务、数字化教育等)的数据格式或数据处理系统互操作。另一方面，我们在很大程度上仍然受印刷载体影响，习惯将信息单元定义为具有固定结构、内容、载体形态和处理方式的单一化永久性集合，而不是定义和组织为可动态地变化、传递、转换、抽取和集成的信息集合，从而使我们对动态、开放的信息组织方式及其标准与技术体系至少是不够敏感、甚至可能有一些潜在的抵触。再一方面，我们在信息处理与服务上也常缺乏一种与其它系统协作和互操作的开放思维和操作机制，缺乏一种基于未来、基于变化的发展态势，习惯于从自己的固有功能和特殊性(其中许多其实是局限性)出发来孤立地考虑和建立相关的技术机制和服务机制。在网络化信息资源迅速增长并日益成为用户的主要信息环境、网络信息系统日益注意信息处理的灵活性和互操作性的今天，这种状况使图书馆系统处于严重不利局面，也使图书情报人员的信息组织与处理能力很难应用到本来是非常相通的其它领域。

我们应该充分认识 XML 技术体系对开放式可扩展智能化信息处理机制的促进作用和对未来信息处理与交换环境的整合作用，充分认识 XML 为我们改造提升图书情报系统能力和参与其它应用领域信息系统所提供的战略机遇，积极探索利用 XML 技术进行信息组织、处理和交换的方法和机制，使我们对这个网络化数字化信息环境信息处理核心技术能做到领先一步、技高一筹，并在此基础上充分开发网络化知识化信息服务的能力，从而开创信息服务与信息系统的新境界。

参考文献：

[1] XML Home Page(http://www.w3.org/XML/)
[2] 张晓林. 基于 XML 的信息组织与处理：1. XML 技术体系. 情报科学，2001(待发)
[3] Robin Cover. The XML Cover Pages (http://www.oasis-open.org/cover/)
[4] XHTML 1.0: The Extensible HyperText Markup Language. A Reformulation of HTML 4 in XML 1.0. W3C Recommendation. Jan 26, 2000 http://www.w3.org/TR/xhtml1/
[5] Scalable Vector Graphics 1.0 Specification. W3C Candidate Recommendation Nov 2, 2000. http://www.w3.org/TR/2000/CR-SVG-20001102/
[6] Synchronized Multimedia Integration Language 1.0 Specification. W3C Recommendation June 15, 1998. http://www.w3.org/TR/REC-smil/
[7] Mathematical Markup Language (MathML) Version 2.0. W3C Recommendation 21 February 2001. http://www.w3.org/TR/MathML2/
[8] Speech Synthesis Markup Language Specification for the Speech Interface Framework. W3C Working Draft Jan 3, 2001. http://www.w3.org/TR/2001/WD-speech-synthesis-20010103/
[9] MusicXML Definition. Early Review Draft: Jan 12, 2001 http://www.musicxml.org/xml.html
[10] Geography Markup Language (GML) v1.0. OGC Document Number: 00-029. 12-May-2000 http://www.opengis.org/techno/specs/00-029/GML.html
[11] CML-XML: Chemical Markup Language. 1. Basic Principles (JCICS 1999) http://www.ch.ic.ac.uk/chimeral/documents/jcics99/jcics99.pdf
[12] 张晓林. 元数据开发应用的标准化框架。现代图书情报技术，2001(待发)
[13] Resource Description Framework (RDF) Model and Syntax Specification. W3C Recommendation 22 February 1999. http://www.w3.org/TR/REC-rdf-syntax/
[14] Guidance on expressing the Dublin Core within the Resource Description Framework (RDF) http://www.ukoln.ac.uk/metadata/resources/dc/datamodel/WD-dc-rdf/
[15] PICS Rating Vocabularies in XML/RDF. W3C NOTE 27 March 2000. http://www.w3.org/TR/rdf-pics
[16] Resource Description Framework. (RDF) Schema Specification 1.0. W3C Candidate Recommendation 27

March 2000. http://www.w3.org/TR/rdf-schema/
[17] XML Topic Maps (XTM) 1.0. TopicMaps.Org Approved Specification http://www.topicmaps.org/xtm/1.0/
[18] Semantic Web Activity http://www.w3.org/sw/
[19] DARPA Agent Mark Up Language (DAML) http://www.daml.org/
[20] Ontology Interchange Language (Ontology Inference Layer) http://www.ontoknowledge.org/oil/
[21] Simple HTML Ontology Extensions. http://www.cs.umd.edu/projects/plus/SHOE/
[22] ebXML Home Page http://www.ebxml.org/
[23] e-GIF Home Page http://www.govtalk.gov.uk/egif/home.html
[24] K. Thibodeau. Building the Archives of the Future, D-Lib Magazine, Feb 2001, 7(2)
http://www.dlib.org/dlib/february01/thibodeau/02thibodeau.html
[25] Health Level Seven XML Patient Record Architecture http://xml.coverpages.org/hl7PRA.html
[26] Ilya Zaslavsky , et al. XML-based Spatial Data Mediation Infrastructure for Global Interoperability. 13-15 March 2000 http://www.npaci.edu/DICE/Pubs/gsdi4-mar00/gsdi_iz.html
[27] SIF: Schools Interoperability Framework.http://www.sifinfo.org/overview.html
[28] ADL Sharable Courseware Object Reference Model (SCORM).
http://www.adlnet.org/Scorm/docs.SCORM_2.pdf
[29] NewsML - Markup for the third millennium http://www.iptc.org/NMLIntro.htm
[30] XML for TEI Lite http://xml.coverpages.org/tei.html
[31] H. Van de Sompel, & C. Lagoze. The Open Archives Initiative Protocol for Metadata Harvesting. Version 1.0, 2001-01-21. http://www.openarchives.org/OAI/openarchivesprotocol.htm [30] NLMMedline DTD revised December 11, 2000 http://www.nlm.nih.gov/databases/dtd/nlmmedline.dtd.
[32] van Horik, R. "Archives and Photographs: the 'European Visual Archive' Project (EVA)", Cultivate Interactive, issue 3, 29 January 2001 URL: http://www.cultivate-int.org/issue3/eva/
[33] NCEAD Guidelines for XML EAD http://scriptorium.lib.duke.edu/ncead/xml-guidelines.html
[34] Making of America II DTD。http://sunsite.berkeley.edu/MOA2/papers/DTD.html
[35] MARC XML http://www.logos.com/marc/marcxml.asp、BiblioML Project.
http://www.culture.fr/BiblioML/en/index.html
[36] Miller, D. R. XML: Libraries'Strategic Opportunity. Library Journal Digital.
http://www.libraryjournal.com/xml.asp
[37] Marcia Lei Zeng. Mapping Metadata Elements of Different Formats. E-Libraries Conference, May 2001.
http://www.infotoday.com/it2001/e-libraries.htm
[38] Li, Ying, et al. Bibliographic data mining: automatically building component part records for e-journal articles on the Internet. Journal of Internet Cataloging. 参见 Medlane XMLMARC http://xmlmarc.stanford.edu/

本文最初发表在《情报科学》2001 年第 9 期第 964-971 页，续 983 页。