Elsevier文档结构规范的分析研究*

陆新民，甘莉，时华
（中国科技出版传媒股份有限公司，北京 100717）

摘要：结构化的文档格式规范是图书和期刊等出版物实现数字化和按需出版、在线发布、全文数据库建设、数据共享的基础和支撑。文章介绍了Elsevier文档结构规范的基本内容，描述了图书和期刊元素结构，并基于Elsevier的文档结构规范完成了对国内期刊出版物的标引试验。实验结果表明Elsevier的DTD规范虽然并不能完全适合中文出版物，但作为领先数字出版公司的、已通过实践生产检验的企业标准，对国内企业甚至数字出版行业建立内容资源结构规范具有重要的参考和借鉴作用。

关键词：Elsevier；内容结构化；DTD；结构化文档；数据规范；XML
中图分类号：G230.7
DOI：10.3772/j.issn.1673-2286.2015.11.007

随着数字技术和信息技术的兴起和发展，传统出版业正向数字出版转型，各种出版物的出版形式、传播手段、阅读方式、市场主体以及商业模式都在不断地发生变化，这些变化深刻的影响着出版物的内容结构。多样化的数字资源是否具有统一的结构属性？对结构属性怎样定义才能达到合适的颗粒度？对于这些问题，国外大型出版机构通过建立本公司的文档结构规范正逐步给出答案[1-3]。部分国外大型出版机构也已完成本企业文档结构规范的制定，并成功应用于企业出版物的数字化生产过程，如Elsevier的文档结构规范、NLM JATS标签集（美国国立医学图书馆的期刊文档标签集）。

Elsevier作为STM（科学、技术、医学）领域世界领先的产品和服务提供商，每年出版超过两千种期刊和近两万种图书，其期刊和图书均采用统一的文档结构规范进行描述。本文对Elsevier的文档结构规范进行了结构分析和标引试验，以期为国内企业和数字出版行业建立出版物文档结构规范提供参考和借鉴。

1 Elsevier文档结构规范分析

随着计算机辅助生产（Computer-Aided Production，CAP）的发展，Elsevier超过两千种STM期刊和越来越多的图书实现了基于XML的数字资源生产。XML用于输出成期刊和图书的纸质印刷版，同时用于Elsevier的数字产品如ScienceDirect平台，从XML中提取出来的摘要则应用于Scopus和PubMed平台。

Elsevier的图书和期刊内容遵循XML优先的原则，所有的文章和图书都转换为XML，作为所有纸质或电子产品输出格式的基础。期刊和图书的XML文件使用Elsevier的文档结构规范进行描述；XML的文件结构使用文件类型定义（DTD）进行定义，Elsevier通过制定DTD系列标准来定义其文档结构规范。1992年，Elsevier制定了第一个版本的DTD用于描述期刊文章全文。此后不断更新完善，从3.0版本到4.1、4.2、4.3版本，目前最新的为5.0版本。

1.1 总体内容结构分析

Elsevier的整套DTD结构图见图1，包括基本集CEP、期刊文章JA、期刊期次SI、图书BK、增强片段EF等DTD模块，同时CEP涵盖了数学标记语言MathML及CALS（Computer-Aided Logistics Support）table的内容[4]。

* 本研究得到国家科技支撑项目“科技文献动态数字出版技术研发与应用示范”（编号：No.2012BAH90F00）资助。

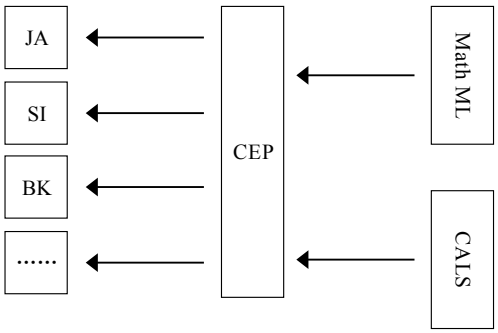

图1 Elsevier的DTD结构图

CEP为整套DTD的基本集，不同类型的出版物描述都是以此为基础的。另外，Elsevier在实践中发现，对于其出版物中的一些复杂公式和表格，尚无法使用MathML和CALS完成描述。因此，Elsevier在CEP中也增加了一些对数学公式描述的标签，同时使用了1个Extend CALS（扩展CALS表格模型），使得整套DTD能够描述Elsevier出版物中的所有公式和表格。对于参考文献，Elsevier也单独定义了结构化参考文献（Structured bibliographic references）。

结构图中各内容的描述如下：

- CEP：Common Element Pool。该套DTD为基本集。
- MathML：Math Markup Language。在描述数学公式上，Elsevier采用了MathML标准。
- CALS table：Computer-Aided Logistics Support（SGML、XML显示表格的标准）。在描述表格上，Elsevier采用了CALS表格模型，并对其进行了扩展。
- JA：Journal Article。该套DTD用于描述期刊文章。
- SI：Serial Issue。该套DTD用于描述期刊期次。
- BK：Book。该套DTD用于描述图书。
- EF：Enhancement Fragment。该套DTD用于描述添加到已在线出版的期刊和图书的内容，如exam元素。

1.2 图书内容结构分析

1.2.1 图书顶点元素

顶点元素可以作为XML文件的根元素。图书顶点元素包括book、introduction、chapter、simple-chapter、examination、fb-non-chapter、glossary、bibliography、index等。各元素描述图书的不同内容，详见表1。

表1 图书顶点元素

顶点元素	描述内容
book	图书的主干
introduction	导论
chapter	图书章节
simple-chapter	单图书章（仅用于PreCAP backfile conversion project）
examination	测验或问答
fb-non-chapter	图书中未按章区分的区域。如foreword（前言），preface（序），about the author（作者简介），back matter（文后部分，如appendices附录）
glossary	词汇表
bibliography	引文或参考文献
index	索引

1.2.2 book元素简介

在Elsevier的DTD中，图书的主干结构存储为一个XML文件。该文件的根元素是book，用于描述图书的主干结构，同时通过ce：include-item调用chapter、index等除book外的顶点元素，构成对整本图书的完整描述。

Book元素包括info、top、front、body（下含volume、part、section等子元素）、rear、ce:floats等子元素，必选的为info、top、body三个子元素。其中，info元素描述图书的基本信息，如DOI、ISBN、版权、主题分类等。top元素描述图书的标题、版权页、声明页内容。图书的文前部分（如序、前言）、正文部分、文后部分（如附录）分别使用front、body、rear元素进行描述。ce:floats作为图书节点下可选的子元素，是图片、表格等内容的容器元素。图书book顶点元素结构图见图2。

1.2.3 调用元素

图书主干结构文件的各元素通过ce：include-item元素调用包括CEP元素集下的其它元素，形成整个图书的XML文件。下文详细列出了book的子元素front、body、rear可调用元素的内容。

front子元素可调用内容包括contributing authors（作者）、reviewers（审稿人）、foreword （前言）、preface（序）、acknowledgement（致谢）、biography

（作者简介）等。这些内容放在fb-non-chapter下，front通过ce:include-item调用fb-non-chapter元素。

body、part、section元素可调用的元素包括chapter（章节）、introduction（导论或介绍）、examination（问答或测验）、bibliography（书目或参考文献）等。

rearpart元素（rear的子元素）可调用的元素包括golssary（词汇表）、bibliography（书目或参考文献）、index（索引）、fb-non-chapter（如附录）等。

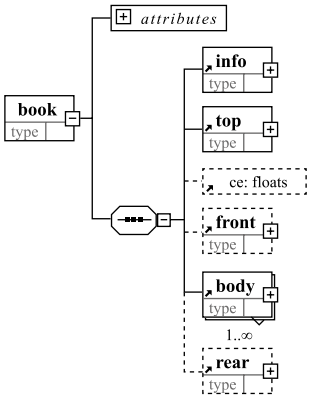

图2 图书顶点元素结构图

1.3 期刊内容结构分析

1.3.1 期刊顶点元素

用于描述期刊的包括两个DTD：Journal Article DTD（JA）和Serial Issue DTD（SI），JA用于描述期刊文章，SI用于描述期刊期次，具体描述如表2所示。

表2 期刊顶点元素

DTD	顶点元素	描述内容
JA	article	期刊的文章
	simple-article	单一文章，如社论（editorials）、讣告等
	book-review	图书评论的文章
	exam	测验式文章（包含问题和答案），如继续医学教育（CME）考试
SI	serial-issue	期刊的期次

1.3.2 serial-issue元素简介

serial-issue元素包括issue-info、issue-data、issue-body三个子元素。issue-info描述期刊期次的唯一标识信息，如DOI号、ISSN号、年卷期等。issue-data描述属于期刊期次的数据，如页码、封面图片等。issue-body提供该刊期与其下属内容之间的关联，是对目录的描述，下含ce：include-item元素或issue-sec元素。期刊期次的元素结构图见图3。

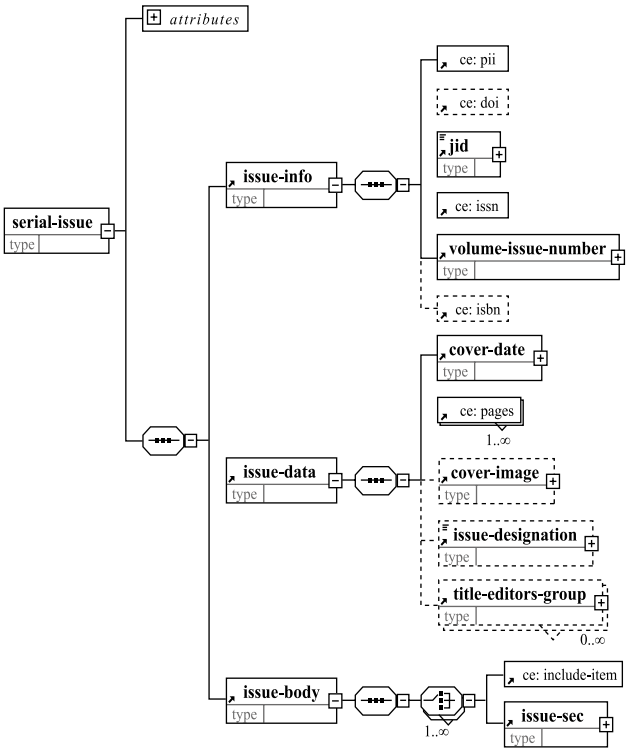

图3 期刊期次的元素结构图

JA下包括四个顶点元素，其中，article、simple-article、book-review三个元素的结构基本一致，下属元素分别描述顶点元素的基本信息（item-info）、文前信息（head）、正文信息（body）、文后信息（tail）以及图表的信息（ce:floats）。exam元素与上述三个元素的结构不太相同，没有描述正文信息和文后信息的元素，取而代之的是描述测验的问题和答案的元素，分别为ce:exam-questions、ce:exam-answers元素，这两个元素可以重复多次出现。

article的子元素item-info描述文章的基本信息，记录期刊和文章在Elsevier系统里的编号、文章的pii和doi号等。head子元素描述文章的标题、作者、关键词、摘要以及时间信息等内容。body子元素下含必选元素ce:sections，用于描述文章的段落内容，其他四个元素均为可选元素。文章顶点元素的结构图见图4。

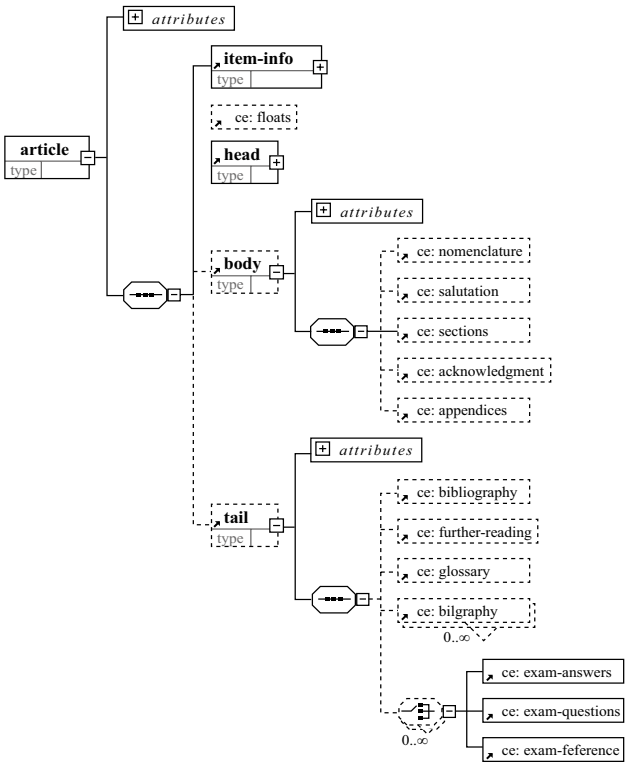

图4 文章顶点元素的结构图

2 国内出版物标引试验

经过5.1版本到5.4版本的不断改进，Elsevier的文档结构规范已经较为完善。规范中的DTD结构清晰、元素详尽。通过分析图书和期刊的顶点元素及其调用的元素可知，这些元素覆盖了一本图书或期刊文章的大部分内容。

Elsevier作为荷兰的图书出版集团，其编写的文档结构规范更适用于英文版的图书和期刊。因此在用于中文出版物标引过程中，要结合中文科技类期刊的特点，对Elsevier的文档结构规范进行扩展，以适用于中文出版物的内容结构，并为制定企业级文档结构规范打下良好基础。

为此，选取《软件学报》的一篇期刊文章《利用块几何约束及视差概率的立体匹配算法》进行结构标引试验。对该文章全部内容进行逐一查看、识别，找到Elsevier文档结构规范的相应元素进行标引，形成XML文件。标引文章的示意图见图5。

科技类图书和期刊作为该领域研究成果的展现方式，充分体现了科技类知识的特点。为更好地实现科技类信息的交流和传播，科技类图书和期刊存在着大量的科技符号、表格和图形，对此类内容的标引、结构化存储，是研究科技类图书期刊数字出版的重点方向。下文重点介绍使用Elsevier文档结构规范标引期刊文章的公式、表格和图片。标引文章图片和公式示意图见图6。

2.1 公式

数学公式是科技类图书和期刊中较常见的公式类型。数学公式表达方式对科技期刊文章被引用有极大

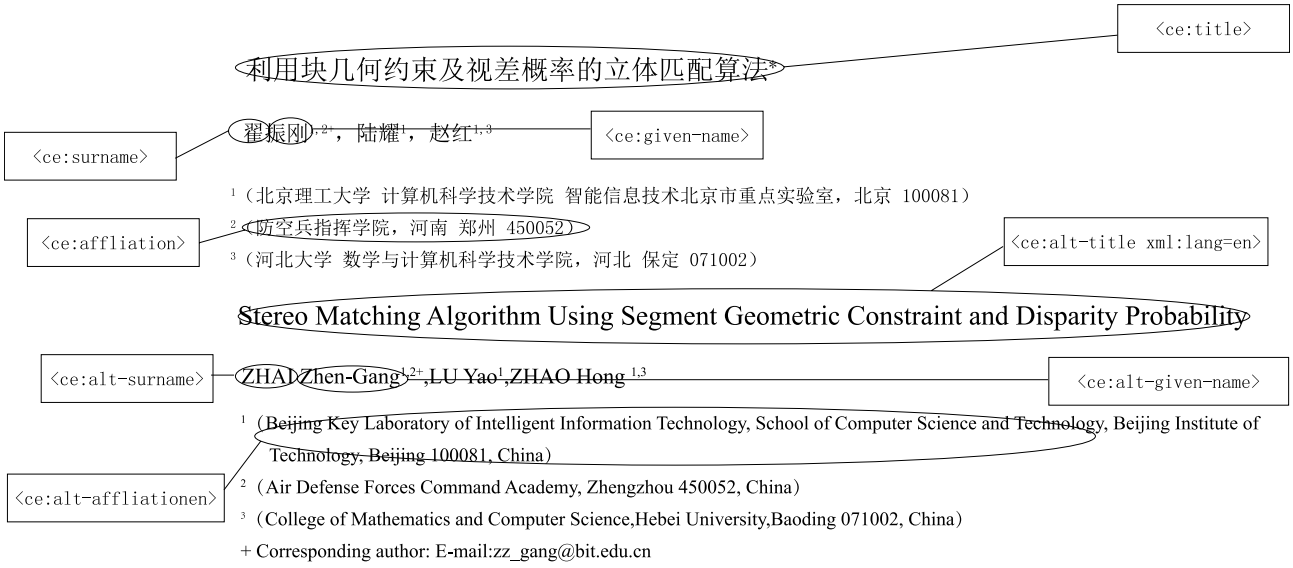

图5 标引文章示意图

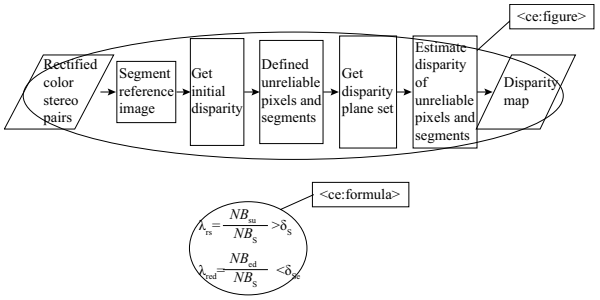

图6 标引文章图片和公式示意图

影响，原因是其表达方式不规范容易导致读者对公式产生恐惧感[5]。只有对数学公式进行规范化的标引和存储，才能在转换和解析过程中正确地进行表达，不至于产生错误或歧义。MathML作为国际通用的数学标记语言，是一种基于XML的标准，用来在互联网上书写数学符号和公式。在编制企业级文档结构规范中，将使用MathML作为描述数学公式的语言。

```
<ce:formula id="f7a"><ce:label>(7a)</ce:label>
<mml:math altimg="si56.gif">
  <mml:mi>&alpha;</mml:mi>
  <mml:mo>=</mml:mo>
  <mml:mo>&int;</mml:mo>
  <mml:mfrac>
    <mml:mrow>
      <mml:msup>
        <mml:mi mathvariant="normal">d</mml:mi>
        <mml:mn>3</mml:mn>
      </mml:msup>
      <mml:mi>k</mml:mi>
    </mml:mrow>
    <mml:msup>
      <mml:mrow>
        <mml:mo>(</mml:mo>
```

图7 使用MathML表示的数学公式

MathML由两种基本独立的标记组成：一种是表现型标记（Presentation Markup），用来描述数学公式的层次结构；另一种是内容型标记（Content Markup），用来描述数学公式的逻辑内容。Elsevier的DTD倾向于使用表现型标记，希望数学公式是由数学软件生成，而不是由人工编写的。当按照规范的流程生成数学公式时，文章将会包含表现型的数学标记。表现型标记可以精确地控制一个数学公式的外观，如在网页上的显示，或在打印纸上的打印样式等。

为了实现向后兼容性，Elsevier的文章和图书目前不使用MathML版本2中的部分元素和属性，Elsevier的DTD规范中详细列出了这些内容。但是，Elsevier亦表示，随着时间的推移和认知的不断变化，部分不使用的元素如mml:maction，未来也可能再次被使用。图7展示了使用MathML表示的数学公式。

2.2 表格

表格同样是科技类图书和期刊中较常见的内容，Elsevier对CALS表格模型进行扩展来完成对Elsevier文章和图书中表格的描述。CALS表格模型作为SGML/XML表示表格的事实标准，由OASIS（结构信息标准化促进组织）发布，OASIS严格审查了CALS表格模型及支持它的软件，以此形成了OASIS交换模型[6]。

Elsevier在实践中发现，尽管CALS表格模型定义了大量的参数，但仍然无法满足Elsevier的文章和图书中表格的描述需求。因此，Elsevier的DTD扩展了CALS表格模型中的border元素，并就column描述进行了完善。图8展示了使用CALS表格模型的表格。

```
<ce:table id="tbl001" frame="topbot" colsep="0" rowsep="0">
  <ce:label>Table 1</ce:label>
  <ce:caption id="c4">
    <ce:simple-para id="sp4">Sm-Nd data.</ce:simple-para>
  </ce:caption>
  <tgroup cols="6">
    <colspec colname="col1"/>
    <colspec colname="col2"/>
    <colspec colname="col3"/>
    <colspec colname="col4"/>
    <colspec colname="col5"/>
    <tb:colspec colname="col6"/>
    <thead>
      <row valign="top" rowsep="1">
        <entry namest="col1" nameend="col2">Eclogites</entry>
        <entry>Sm</entry>
        <entry>Nd</entry>
        <entry><ce:sup loc="pre">147</ce:sup>Sm
              / <ce:sup loc="pre">144</ce:sup>Nd</entry>
        <entry>Yield (%)</entry>
      </row>
    </thead>
    <tbody>
      <row valign="top">
        <entry>162a</entry>
```

图8 使用CALS表格模型的表格

2.3 图片

Elsevier的DTD实现了对三种不同类型图片的结构化描述，其中，行内图片在ce:display元素内描述，跨列图片在ce:floats元素内进行描述，摘要内的图片则在ce:abstract元素内描述。这三个元素的子元素ce:figure用于描述实际图片的信息，而ce:figure的子元素ce:link实现对图片资源的调用。

对于图片资源或者其它外部资源文件（如mp3、video

等），Elsevier是在XML中采用嵌入外部实体链接的方式完成的（使用ce:link元素）。目前Elsevier定义的资源类型有：TEXT（纯文本文件）、IMAGE（GIF、JPEG、TIF格式文件）、VIDEO（AVI、MP4、MPEG格式文件）、APPLICATION（其它应用程序文件、脚本、可执行文件）、XML（外部XML文件，如矢量图或者化学式等）。

3 Elsevier文档结构用于中文文档的建议

通过利用Elsevier文档结构规范标引国内出版物的试验，总结出Elsevier应用于中文出版物的改进建议。

3.1 语言

在Elsevier的DTD中，语言属性xml:lang只支持de|en|es|fr|it|pt|ru七种语言，并未包括中文zh。对于中国的大多数期刊来说，语言以中文为主，只有少部分为纯英文期刊。在标引过程中，将DTD语言的默认值调整为中文。同时，对于有中英文两种语言的内容，如作者、摘要、关键词，通过语言属性xml:lang进行区分。

3.2 具有中国特色的内容

除了语言属性外，国内出版的期刊和文章的部分内容，如期刊CN号、主管单位、中图法分类号等具有中国国情的信息，在Elsevier的期刊DTD中未进行描述。在设计企业文档结构规范时，需增加元素描述该部分信息。

3.3 增加的其他元素

由于DTD的内容众多，在实际生产过程中，可分别由作者、编辑、专家对各自擅长的元素分别进行标引，确保各项标引内容的准确性，提高数据生产加工的效率和质量。在设计企业级的DTD时，可增加属性描述该信息。同时，文档结构规范的设计还需考虑到系统间数据传输需要存储的一些字段，如碎片化文件的大小、MD5码、文件名等，需要扩展新的元素进行描述。

4 结语

本文对Elsevier图书和期刊的文档结构规范进行了详细分析，并对国内期刊出版物进行了标引实验。Elsevier的文档结构规范虽然并不能完全适合中文出版物的文字和内容，但作为领先数字出版公司的、已通过实践生产检验的企业标准，对国内企业甚至数字出版行业建立内容资源结构规范具有重要的参考和借鉴作用[7]。为确保编制的规范成功应用于企业的出版物，今后将在Elsevier文档结构规范分析的基础上，制定适合于本企业出版物的企业文档机构规范，并使用测试检验工具完成对文档结构规范的验证，并将研究成果应用于内容结构化标引与拆分系统的研制。

参考文献

[1] 沈锡宾,李鹏,王红剑,等. 中华医学会系列期刊全文电子文档交换和存储标准初探[J]. 中国科技期刊研究,2015,16(5):475-479.

[2] 白杰,杨爱臣. XML结构化数字出版的特点与流程[J]. 出版广角,2015(05):28-31.

[3] 沈锡宾,顾佳,包靖玲,等. 中国科技期刊文档格式标准化任重道远[J]. 编辑学报,2013,25(1):27-30.

[4] ELSEVIER[EB/OL].[2015-05-20].http://www.elsevier.com/wps/find/authorsview.authors/dtds_htm.

[5] 谢文亮,张宜军. 科技期刊中数学公式的规范表达[J]. 编辑学报,2013,25(3):240-242.

[6] CALS_Table_Model[EB/OL].[2015-05-29]. https://en.wikipedia.org/wiki/CALS_Table_Model.

[7] 刘冰,游苏宁. 我国科技期刊应尽快实现基于结构化排版的生产流程再造[J]. 编辑学报,2010,22(3):262-266.

作者简介

陆新民，男，1971年生，硕士，中国科技出版传媒股份有限公司副编审，研究方向：信号与信息处理，E-mail：luxinmin@mail.sciencep.com。
甘莉，女，1981年生，硕士，中国科技出版传媒股份有限公司编辑，研究方向：数字出版，E-mail：ganli@mail.sciencep.com。
时华，男，1981年生，硕士，中国科技出版传媒股份有限公司高工，研究方向：项目管理、数字出版，E-mail：shihua@mail.sciencep.com。

Analysis and Study on the Structural Standardization of Elsevier's Documentation

LU XinMin, GAN Li, SHI Hua
(China Science Publishing & Media Ltd., Beijing 100717, China)

Abstract: Standardization of structuralized documentation format is the basis and support for publications such as books and journals to be digitalized, printed on demand, released online, data-banked in their entirety and to enjoy data sharing. This paper presents the fundamentals of the structural standardization of Elsevier's documentation and describes in detail the elemental structures of books and journals, and based on the standardization of Elsevier's documentation structure, the indexing experiment on domestic journal publications has been completed. Although Elsevier's DTD Standardization cannot completely meet the characteristics of the language and content of the Chinese publications, this corporate standardization, created by the leading digital publishing corporation and having gone through the test of practice and production, can still be learned from or used as an important reference in setting the structural standardization of content resources for the whole digital publishing industry as well as this company.

Keywords: Elsevier; Content Structuralization; DTD; Structuralized Documentation; Data Standardization; XML

（收稿日期：2015-09-09）