图书表示模型BITS及其对中文科技图书的适用性分析*

王晓光，刘怡丹
（武汉大学信息管理学院，武汉 430072）

摘要：出版物的结构化处理是语义出版发展的基础。图书交换标签集（Book Interchange Tag Suite，BITS）是一套面向科技类图书的结构化表示模型。本文通过分析BITS模型的元素和结构，比较BITS与期刊标签集JATS以及电子书标准Docbook和DITA的异同，并结合中文科技类图书的特点，以探讨BITS对中文图书的适用性和改进方向。

关键词：数字出版；图书；内容结构化；BITS；JATS
中图分类号：G250.73
DOI：10.3772/j.issn.1673-2286.2016.8.006

1 引言

出版物的结构化处理是实现语义出版的必经环节。经过结构化处理后的出版物多以XML语言表示，内涵丰富的树状结构语义标签，体现内容的有序层次化知识结构，有利于内容的存档、重用、格式转换与互操作，能够更好地满足出版物内容发布和再利用需求。

图书交换标签集（Book Interchange Tag Suite，BITS）是由美国国立医学图书馆下属国立生物技术信息中心开发的一套基于XML的图书表示模型，旨在为科学、技术、医学（Scientific，Technical and Medical，STM）出版提供一套通用的图书数据存储和交换格式，便于图书内容资源的表示、存储和多样化开发利用。BITS问世于2013年12月，之后经过不断地发展和修正，于2016年2月更新为BITS 2.0。从本质上看，BITS是一套基于XML的描述图书叙事化内容及其元数据的模型，不仅适用于单本图书，还适用于系列图书和成套图书，可用于出版商之间、出版商与发行商之间进行内容的交互、存档和格式转换[1]。源于出版物内容与结构的相似性，BITS的应用对象不仅是STM图书，还包括政府报告、参考书、系列丛书、会议论文集和百科全书等多种形式的出版物[2]。

目前，越来越多的出版社和学者开始关注和应用BITS，并在其基础上开发出多个更具领域性特征的个性化版本，如剑桥大学出版社在BITS基础上开发的图书内容存档标准CUP-BITS[3]；Wheeles在Silverchair信息系统中对BITS的应用[4]；加拿大学者门户网站（Scholars Portal）基于BITS制定的电子书格式标准管理超过60万本电子书[5]。

相较于西方专业出版机构，国内出版社对科技类图书的结构化处理工作起步较晚，但发展速度很快。近五年，很多出版机构在借鉴Docbook、DITA等电子书模型基础上，提出自己的图书结构化标准与规范。为进一步提高中文STM图书结构化处理和编辑的规范性，提高内容资源的互操作效率，有必要分析BITS的元素与结构特点及其对中文科技类图书的适用性。

2 BITS元素、标签与结构

2.1 BITS的元素与标签

图书是内容的集合，内容单元可以进行粗细粒度不

* 本研究得到中组部“青年拔尖人才”支持计划和教育部“新世纪优秀人才”支持计划资助。

同的划分。较粗粒度层面上，图书不仅包括“正文本”，即图书的各章节内容，还包括各种具有特殊意义的“副文本”，如前言、后记、目录、版权页等；较细粒度层面上，图书的内容单元包括标题、段落、公式、表格、句子，甚至词汇。图书内容单元的类型及语义内涵是图书模型（Book Model）中元素（element）及其对应标签（tag）的设计依据。

一般说来，根据图书内容单元的功能和意义，可以把图书内容单元对应的元素和标签分为三种类型，分别是结构类、内容类和表示类[6]。结构类标签指架构图书“骨架”的标签，是封装内容类元素的容器，界定图书的结构，并没有实证性意义的标签，如代表图书前置部分的标签、后置部分的标签、图书主体部分的标签等；内容类标签是用于标记填充“骨架”的“血肉”，这类标签指明元素的具体语义，如代表问题部分的标签、答案部分的标签等；表示类标签是用于图书的视觉表现标记，包括字体、字号、下划线、上标、下标等，如表示缩写、斜体的标签等。

BITS模型包含309个元素及其对应的标签，158个属性，基本实现对图书构成要素及其组成结构的清晰、完整和规范的描述。按照上述分类体系对BITS模型中的标签划分后的示例如表1所示。

表 1 BITS标签的分类示例

标签类型	标签示例
结构类标签	<front>前置部分；<back>后置部分；<body>主体部分；<book-app>图书附录； <book-app-group>图书附录集；<sec>小节；<float>浮动部分
内容类标签	<question>问题；<answer>答案；<answer-set>答案集；<array>列表；<article-title>文章标题； <chapter-title>章标题；<city>城市；<code>编码；<conf-date>会议日期；<disp-formula>方程式； <era>年代；<event>事件；<speaker>发言人；<speech>演讲；<state>声明
表示类标签	<abbrev>缩写；<bold>黑体；<disp-quote>显示引用；<italic>斜体；<monospace>单间隔；<overline>上划线； <roman>罗马字体；<sans-serif>无衬线字体；<sub>下标；<sup>上标；<underline>下划线

顶层元素是XML文件的根元素。在BITS中根元素有两个，分别为<book>和<book-part-wrapper>。其中<book>包含一整本图书，例如一本教材或者专著；<book-part-wrapper>包含图书的一个独立的部分，例如一章或者一节。

2.1.1 <book>元素

顶层元素<book>可包含五个部分，分别为（1）集合元数据（<collection-meta>，可选部分），用于描述当前图书所在丛书的元数据；（2）图书元数据（<book-meta>，可选部分），描述当前图书的元数据，包括书名、图书出版日期、出版者、版权声明等；（3）前置内容（<front-matter>，可选部分），包含正文前的扉页内容，例如题词、前言、序言等；（4）主体部分（<book-body>，可选部分），指图书的正文内容，包括文本和图像等；（5）后置内容（<book-back>，可选部分），包含术语表、附录、参考文献列表等，附录也包括浮动元素<float-group>，例如图表、数据和图书的侧边栏内容。<book>元素的结果及各元素间的包含关系如图1所示[7]。

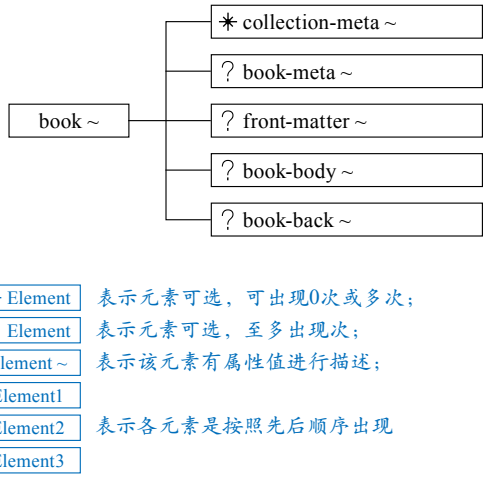

图 1 <book>元素的结构

2.1.2 <book-part-wrapper>元素

图书片段元素<book-part-wrapper>主要用于表示图书内部的章节性模块。该元素所定义的模块在内容

和逻辑上具有相对的独立性[7]，其内部可包含的部分如图2所示。

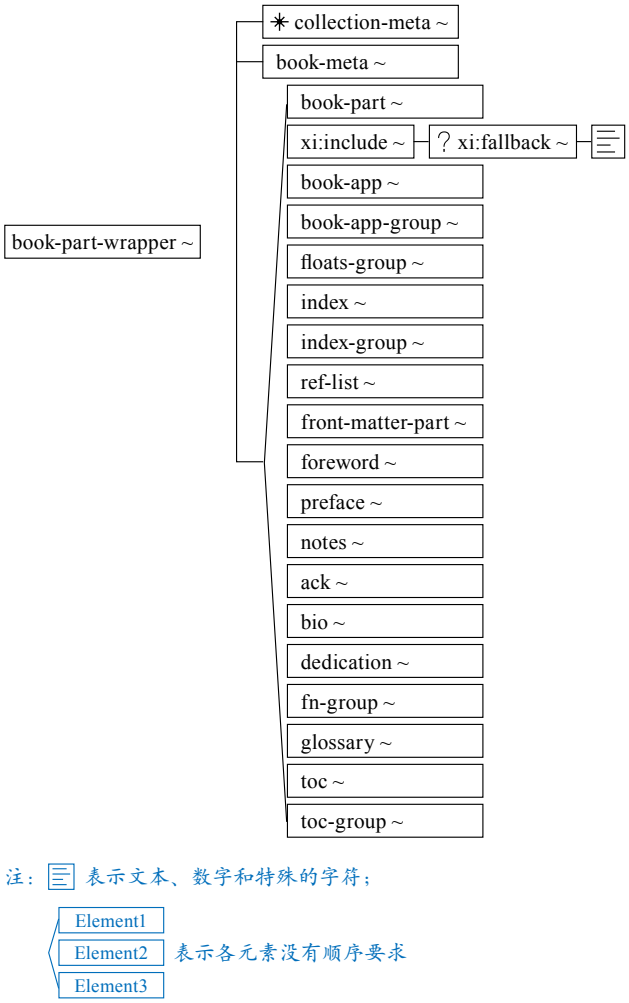

图 2 <book-part-wrapper>元素的结构

<book-part-wrapper>元素可包含的子元素除了与元数据相关的集合元数据（<collection-meta>）和图书元数据（<book-meta>）外，还有19个可选的叙事性内容元素，如表2所示。

2.2 BITS的结构

BITS模型结构代表BITS对图书的理解，反映图书中各内容单元组合关系和模块特征。在SGML和XML产生前，人们对图书的理解往往考虑其物理结构和编辑体系结构，由此形成封面、页面、页眉、页脚、边白等概念，这种分类体系与印刷图书的物理实体性特征有很大关系。进入数字出版时代后，电子图书的兴起使得以物理图书为基础的概念体系产生不足，如何从认识论的角度重新认识图书的构成，成为图书模型创新的关键，也是图书脱离印刷纸张继续存在的逻辑基础。

表 2 <book-part-wrapper>中的19个可选元素

内容相关	图书单元<book-part>；前置内容<front-matter-part>；浮动元素组<floats-group>
索引相关	索引<index>；索引集合<index-group>；目录<toc>；目录组<toc-group>
附录相关	图书附录<book-app>；图书模块附录<book-app-group>；致谢<ack>；传记<bio>；致辞<dedication>；前言<foreword>；序言<preface>
注释相关	脚注集合<fn-group>；术语表<glossary>；注释<notes>；参考文献列表<ref-list>
其他	独立标记<xi:include>（方便在标记过程中进行独立标记，并且纳入最终整体文档）

从本体角度看，图书作为一种文本形式的内容，本质上存在一种“内容对象的有序层次结构”，即图书内容存在一种有序的、层次化的结构，这种结构基于文本而存在，无论印刷载体还是电子载体，这种结构都长久存在，是图书内容的本体结构[8]。在XML语言产生后，这种本体结构可以使用XML表现出来；从形式上，以XML语言表示的图书呈现为一种树状结构。图书可以分为章，章下面可以分为节，节下面可以分为段落，段落内包括文字、图形、公式等，这种结构对应到XML文档上就是元素及其子元素的嵌套结构。这种树状结构的设计为图书结构化处理和标注实践工作奠定了理论基础。利用BITS对图书进行结构化处理的示例[9]见图3。

3 BITS与其他出版物表示模型的比较

3.1 BITS与JATS

期刊文档标签集/套件（Journal Article Tag Suite，JATS）是由NCBI开发的主要用于科技期刊表示的模型[10]。BITS与JATS是一种继承与扩展的关系，BITS继承了JATS的绝大部分功能，包括多语言环境、扩展性、开放性等。根据图书的特点，BITS在JATS基础上做了元素和结构的扩展。JATS与BITS的比较如表3所示。

```
<book dtd-version="2.0">
  <book-meta>...</book-meta>
  <front-matter>...</front-matter>
  <book-body>
    <book-part id="bid.2" book-part-type="chapter">
      <book-part-meta>...</book-part-meta>
        <body>
          <sec id="bid.3">
            <title>History</title>
            <p>Initially, GenBank was built and maintained at Los Alamos National Laboratory...</p>
          </sec>
          <sec id="bid.4">
            <title>International Collaboration</title>
            <p>In the mid-1990s, the GenBank database became part of the International Nucleotide ...</p>
          </sec>
        </body>
        <back>...</back>
    </book-part>
    <book-part>...</book-part>
    <book-part>...</book-part>
    <book-part>...</book-part>
  </book-body>
</book>
```

图 3 利用BITS表示的图书结构示例

表 3 JATS与BITS对比

		JATS	BITS
适用范围		STM期刊	STM图书
文档构成	前置部分	√	√
	主体部分	√	√
	后置部分	√	√
	浮动部分	√	√
	问答模块	×	√
	评论模块	√	×
组件数量/个	元素	254	309
	属性	135	158

从表3可见，相对于期刊，图书的内容和结构更为复杂，所以需要更多的元素和属性辅助表示。BITS新增的主要元素包括4个。

（1）目录元素（<toc>）。由于期刊文章是单篇，而图书由多个章节组成，丛书由多册单本组成，需要目录来呈现整本图书的内容及章节结构。BITS新增了目录元素（<toc>），用于对图书的目录进行表示。

（2）索引元素（<index>）。一本书的内容较多，篇幅较长，需要索引以方便查阅。BITS新增了索引元素（<index>），对图书的索引进行标注。

（3）问答元素[2]。包含<question>、<question-wrap>、<answer>、<answer-set>、<explanation>等一系列新元素。该问答结构定义的是问题和回答的模式，尽管没有定义测验的形式，但是经过修改，能够按照需求定义问答形式。

（4）更深入的嵌套结构元素。图书中拥有相互嵌套的结构，例如卷、章、节等，JATS结构无法满足。

3.2 BITS与Docbook、DITA

除BITS外，其他知名的图书表示模型还有DocBook和DITA。DocBook是一种主要用于技术文件的标记语言，它使用XML定义了一系列文档元素，是一种将文件内容与样式分开处理的文件规范[11]；达尔文信息分类体系构架（Darwin Information Typing Architecture，DITA）是一种面向主题的DTD规范[12]。BITS与DocBook、DITA三者都专注于交付技术类信息，但是各有所长，同时在实际应用中也有各自的限制。

（1）适用对象。DITA适用于对格式有严格限定的技术手册类出版物；DocBook适用于一般性出版物，文档易于组织和排版；BITS适用于STM领域多种类型的出版物。

（2）基本结构。DocBook、BITS与DITA的基本结构对比如图4所示。DocBook的主体内容以章节（Section）为组织单元，易于整体内容的组织和排版，但无法做到更加细粒度内容的语义表示；对于内容需要频繁修改的文档排版，DocBook也显得力不从心。DITA使用主题（Topic）作为单元组织文档，支持模块化的内容创建；同时允许作者组织并描述特定领域；在生成多种文档格式的内容重用过程中，能够保持内容的高度一致性。BITS的设计完全考虑了图书的有序层级结构，在元素粒度及其属性和关系设计上，BITS考虑得也十分细致和全面。

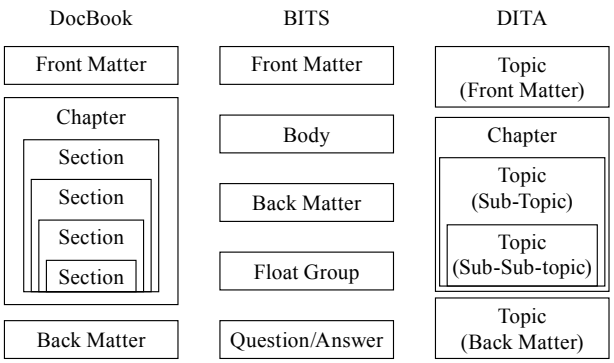

图 4 DocBook、BITS与DITA的结构对比

3.3 BITS的特点

通过分析BITS的元素及其属性，对BITS与DocBook、DITA的比较，可以发现BITS在图书模型表示上具有明显优势，具体表现为三个方面。

（1）适用范围广、内容集成便利。BITS本身是在JATS基础上开发完成的，二者使用的元素标签保持高度一致性。随着两个模型的更新，二者的互操作能力有了很大的保障，这使得期刊和图书的内容集成有了很多便利。

（2）粒度细致，结构灵活。与其他图书模型相比，BITS的元素更为丰富，这使得BITS在内容单元的表示和描述上更加细致，有利于内容单元的提取、集成、检索和再利用。由于使用更为准确和细致的标签，计算机在“理解”内容单元的能力上也得到提升，为内容资源的重组和互联等开发提供便利。

（3）应用方便、学习成本低。DITA包含众多语法和标签定义，在应用过程中学习曲线较陡，在文档生成过程中一般需要Ant命令进行编译；DocBook门槛较低，只需了解基本知识就可以使用；BITS的基础元素和属性数量相对较少，学习难度较低，但是扩展性和灵活性强，便于用户的扩展应用。目前，使用JATS的机构较多，掌握JATS使用方法可以迅速接受BITS。

4 BITS对中文科技图书的适用性

4.1 语言适用性

BITS沿袭了JATS的强大功能，支持包括中文在内的多种语言，“zh”标签代表中文。在BITS中，语言属性@xml:lang用于解释所定义内容的语言类型，所定义的值具有延续性。例如，@xml:lang定义一段内容为中文，则下述段落也默认为中文。若改变语言类型，需对专属语言属性重新定义。此外，大多数语言有不同的书写规范及书写字体，就中文来说，有简体、繁体之分。BITS在支持多种语言的同时，对语言的呈现形式及规范也纳入考虑。

在BITS中，通过语言标签与字体标签进行组合，可以实现不同的中文表现形式。例如：xml:lang="zh-Hant"表示中文繁体书写体；xml:lang="zh-Hans"表示中文简体书写体；xml:lang="zh-Hans-CN"表示在中国大陆使用的中文简体书写体。

对于中国大多数科技类图书来说，语言以中文为主，少部分为纯英文或其他语种。BITS的多语言功能，能够较为全面地匹配中文科技图书在结构化工作中对语言的需求。

4.2 元素适用性

中文科技图书既包含文档层次，也包括页面层次的逻辑元素。在文档层次上，中文科技图书的逻辑元素有封面、书名页、版权页、目录页、前言、正文页、参考文献、附录、术语表、索引等；在页面层次上，中文科技图书的逻辑元素主要包括书名、作者、目录条目、章、节标题、正文段落、图片、表格及其标题、页眉、页脚、页码、注释、索引条目等。逻辑元素的适用性是指文档中的逻辑元素能够被识别[13]，综合文档层次与页面层次的逻辑元素，可以看出，对中文科技图书进行结构化标注，需要识别的主要逻辑元素有封面、书名页、版权页等23个逻辑元素，BITS元素具体适配情况如表4所示。

BITS能够满足中文科技图书主要逻辑单元的调用，图书中主要逻辑元素在BITS中至少能找到一个元素进行对应，并且BITS拥有309个元素能够满足对中文科技图书标注更加细粒度的要求。

4.3 结构适用性

一般说来，中文科技类图书与英文科技类图书的结构无太大差异。除元数据部分外，图书的叙事性内容部分都由三部分构成，分别为前项部分、正文章节部分和后项部分。这三部分与BITS<book>元素的子元素结构一一对应。如BITS<front>元素对应图书的前项部分，<body>元素对应图书的正文章节部分，<back>元素对应图书的后项部分。除这种基本层次结构的对应外，中文科技类图书的内部关联结构，如不同内容单元间的索引关系、引用关系、参照关系等都能利用BITS的元素及其属性关系表示，说明BITS对中文科技类图书有极强的适用性。

4.4 BITS在中文科技图书应用上的不足

BITS虽然对中文科技类图书有极强的使用性，但囿于语言习惯和文化差异，BITS在两个方面还需改进以更加适用中文科技类图书。

表 4 中文科技图书逻辑元素与BITS元素的适配

中文科技图书逻辑单元		BITS中是否有可调用元素	可调用的BITS元素
文档层次	封面	√	<fpage>、<front-matter>、<front-matter-part>
	书名页	√	<book-Title>
	版权页	√	<copyright-holder>、<copyright-statement>、<copyright-year>
	目录页	√	<toc>
	前言	√	<foreword>
	正文页	√	<book-body>
	参考文献	√	<ref>、<ref-list>
	附录	√	<book-app>
	术语表	√	<glossary>
	索引	√	<index>
页面层次	书名	√	<book-Title>
	作者	√	<attrib>
	目录条目	√	<toc-div>、<toc-entry>、<toc-group>、<toc-title-group>
	章标题	√	<chapter-title>
	节标题	√	<part-title>
	正文段落	√	<p>
	图表	√	<graphic>、<fig>
	表格及其标题	√	<table>、<table-wrap>、<table-wrap-foot>、<table-wrap-group>、<tbody>、<tfoot>
	页眉	√	<floats-group>
	页脚	√	<fn>、<fn-group>
	页码	√	<page-range>
	注释	√	<note>、<notes>
	索引条目	√	<index-div>、<index-entry>、<index-title-group>

（1）增加一定的表示类标签。BITS使用下划线<underline>标签表示重点图书中的重点内容，但是在中文图书中，重点内容通常用打黑点的方式表示。此外，部分字词可能会需要增加拼音显示。这些特殊的表现样式要求，在具体工作中虽可以使用CSS定义解决，但增加一种标签更为方便。

（2）增加代表中国特色的元素和标签。中文图书在版权页通常有代表中国国情的信息，如中国版本图书馆CIP数据，它们在BITS中没有对应的元素和标签。针对这种情况，可以增加新的元素对BITS进行扩展实现。

5 总结

本文详细分析图书表示模型BITS的元素、结构及其特点，并将其与JATS、DocBook、DITA进行比较研究，进而论述BITS对中文科技类图书的适用性。研究表明：BITS比DITA和DocBook拥有更强大的图书结构与单元表示能力，能够满足绝大多数科技类图书的结构化处理要求，是一套理想的图书结构化表示模型。需要注意的是BITS并不适用于包含大量图片、以视觉设计为重心的科技类图书，如各种建筑图像集、工程绘图作品等。

总的来说，BITS标准的推出有利于提高图书结构化处理的规范性，而加强图书结构化处理不仅有利于提高图书内容的重用性，也有利于增强图书内容的细粒度检索、集成和重组能力，是语义出版系统发展过程中必不可少的环节。

参考文献

[1] 包靖玲,霍永丰,顾佳,等.美国国立医学图书馆期刊文档标签集概述[J].中国科技期刊研究,2013,24(4):624-627.

[2] BECK J.What JATS users should know about the book interchange tag suite (BITS) [EB/OL]. [2016-06-26].http://www.ncbi.nlm.nih.gov/books/NBK159737/.

[3] EDEN M,CLEGHORN T.An implementation of BITS:the cambridge university press experience[EB/OL].[2016-06-03].http://www.ncbi.nlm.nih.gov/books/NBK350535/.

[4] WHEELES D.Using BITS for non-standard content [EB/OL].[2016-06-03].http://www.ncbi.nlm.nih.gov/books/NBK279829/.

[5] ZHAO W,DAVID R H,KHWAJA S,et al.JATS for ejournals and BITS for ebooks—adopting BITS for scholars portal ebook repository[EB/OL].[2016-05-26].http://www.ncbi.nlm.nih.gov/books/NBK280069/.

[6] 陈孝禹.科技类电子书的内容结构化标注研究[D].武汉:武汉大学信息管理学院,2013.

[7] Book interchange tag suite (BITS) version 2.0 tag library[EB/OL].[2016-03-20].http://jats.nlm.nih.gov/extensions/bits/tag-library/2.0/index.html.

[8] 德罗斯,杜兰德,米洛纳斯,等.文本到底是什么?[J].出版科学,2016(3):5-13.

[9] Body of a book part[EB/OL].[2016-03-16].http://jats.nlm.nih.gov/extensions/bits/tag-library/2.0/element/body.html.

[10] 康宏宇,侯震,李姣.基于JATS数据标准的全文文献管理[J].中国科技期刊研究,2015,26(11):1171-1175.

[11] Hello and welcome![EB/OL].[2016-05-17].http://www.docbook.org.

[12] 高昂,刘钰,邢立强.DITA数字出版技术[M].北京:电子工业出版社,2013.

[13] 陈国光,丁晓青,彭良瑞.一个基于规则的图书逻辑结构提取算法[J].计算机工程与应用,2002(19):53-57,143.

作者简介

王晓光，男，1978年生，教授、博士生导师，研究方向：语义出版、知识组织、数字资产管理、数字人文，E-mail: whu_wxg@126.com。
刘怡丹，女，1993年生，硕士研究生，研究方向：数字出版、数字资产管理。

Research on the Structures of BITS and its Suitability to Chinese STM Books

WANG XiaoGuang, LIU YiDan
(School of Information Management, Wuhan University, Wuhan 430072, China)

Abstract: The structuration of publication is the foundation of semantic publishing. BITS is a new standard for the structuration of STM books. In this paper, the schemas, elements, and structures of BITS are explained. The similarities between BITS and JATS as well as BITS and Docbook, DITA are also analyzed. The suitability of BITS to Chinese STM books is discussed from three aspects, language, elements and structure. The results show that BITS suits the Chinese STM books well, but still needs to be improved according to the specific features of Chinese STM books.

Keywords: Digital Publishing; Book; Book Structuralization; BITS; JATS

（收稿日期：2016-07-11）