

期刊论文元素标引规范研究*

□ 吴琳 / 中国科学技术信息研究所 北京 100038

程煜华 / 北京万方数据股份有限公司 北京 100038

摘要: 文章根据期刊论文自身的特点和应用需求,分析期刊论文元素标引的意义,设计期刊论文元素标引规范的制定流程,并深入分析和研究标引规范的内容结构、元素构成及互操作框架,在此基础上提出了期刊论文的核心元数据、文档结构、深度内容三个方面的标引元素,同时指出了下一阶段研究工作的主要方向。

关键词: 期刊论文, 元素标引, 标引规范, 元数据标引

DOI: 10.3772/j.issn.1673-2286.2011.03.012

数字化的期刊论文以其丰富的学术信息资源改变了学术期刊论文的资源格局。然而,海量的期刊论文信息资源仓储仍是以异构方式存在的,它们提供各自不同的信息存储、组织、检索与浏览方法,使自己的产品在使用方式、字段、接口设计等方面造成差异,不能实现跨仓储的统一查找利用。

为了克服海量文献信息给人们的检索、利用带来的困难,使文献信息更好地为人们所利用,长期以来人们都在探讨文献发展变化的规律,以便更有效地控制和利用文献信息。对文献资源进行标引加工,使其信息有序化是控制文献、传递信息的最好方法。期刊论文标引的实质正是为了实现了对期刊论文信息的控制,满足用户对信息的需求,而对期刊论文信息所进行的编辑和加工工作。

1 期刊论文元素标引的意义

数字期刊论文标引的标准化、规范化具有以下三方面的意义:

(1) 元素及其元数据的描述能够为数字期刊建立一种计算机普遍可以理解的框架体系,使数字期刊资源得以透明、公开,进而使得数字期刊信息简单、有效地被其他信息资源仓储利用。

(2) 元素的描述、标识能够融入互联网通用的语言中,被迅速地装载或转换进各种索引数据库,使数字期刊资源具有可扩展和互操作性。

(3) 标引的论文元素能够实现对庞大繁杂的数字期刊论文数据库信息流的有效控制,它使海量信息有序化,便于检索利用。

元素标引规范是对数字期刊论

文的属性特征进行描述和标引的规范化指导文件。规范化的论文标引便于文献的关键内容在数字出版过程中的传递,避免信息丢失,更重要的是,为用户提供多条检索途径,提高检索结果的适用性及全面性。因此,论文元素标引信息的编排是否科学、规范,直接影响到期刊的编辑质量和期刊社会效益的发挥。

本文所研究的元素标引规范是以字处理软件为基础,研发对文章进行编辑、加工、标引的方法和工具所遵循的规范和标准。与现有的元数据标准相比,本研究对论文的图、表、公式、知识点等深度内容元素进行了描述和标引,使得论文中的信息内容得到最大化的揭示。

本研究在字处理软件基础上,开发一系列编辑和标引插件工具,作者使用字处理软件在论文写

* 本论文受科技部科研院所技术开发研究项目专项资金“全XML流程多样性出版服务研究”项目资助。

作的阶段即可将元素标识插入到文章中,将编辑后由字处理软件输出的文件进行分析处理,转换为对应的XML文件,便于数字出版流程中的不同组织、单位和系统之间的接口兼容,同时满足不同终端用户在检索时的便捷性和全面性,发挥多样化数字出版“一次制作,多元发布”的优势。

2 期刊论文元素标引规范的制定流程

2.1 标引规范制定的基本原则

在标引规范的制定过程中,要充分考虑作者、编辑、排版和使用者的需求和论文资源的特性,并在其间作一最佳平衡和组配。规范制定要遵循以下几个基本的指导原则^[1]:

① 以论文服务应用需求为导向。制定标引规范的最终目的是为期刊论文服务系统用户更好和更充分地揭示信息资源,因此论文服务应用需求应作为最终的权衡标准。特别是在结构的设计、元素的增加与取舍、著录规则的制定等方面,要尽可能地从应用的角度出发,为系统用户提供多层次的检索和服务体系。

② 从期刊论文的特点和文献利用方式出发。由于各类数字资源的特性不尽相同,标引深度和广度不同,以及资源的应用和服务模式也大不相同,因此需要根据期刊论文的具体特点和利用方式来制定规范。

③ 互操作性原则。标引数据的互操作性体现在对异构系统间互操作能力的支持,即在我们设计的标引规范下建立的数据,不仅能方

便地为自己建立的各相关应用系统所操作,还应尽可能地为其他组织或机构所建立的应用系统所操作。因此,在设计标引规范时,要非常慎重地考虑与期刊论文现有的相关标准和规范的一致性。

④ 可扩展性原则。由于数字资源类型非常广泛,并不仅限于期刊论文,且各类应用背景更为复杂,标引规范在数字资源的共性描述基础上,定义面向期刊论文特点的个性元素描述,同时需要预留出一些扩展空间,以适用其他资源及其应用,允许扩充一些元素、子元素或属性值。

2.2 标引规范制定的主要环节

期刊论文元素标引规范的制定是基于对期刊论文特征的调查与分析、面向系统应用而制定标准化规则的过程,主要环节如下^[2]:

① 资源分析:分析期刊论文资源的特点和文献利用方式;

② 建立模型:详细列出相关元素,表明其相互之间的关系;

③ 属性提取:从资源利用的角度提出属性要求;

文献调研:从现有标准和相关规范中获得参考和引用;

④ 元素精炼:分析期刊论文资源的特点和文献利用方式,精简关键元素;

⑤ 规范控制:考察需要进行控制的属性;

限定规则:规定属性限定原则(子元素、修饰词、属性值和标引代码);

著录规则:制定详细的著录规则,指导数据标引;

⑥ 置标方案:使用XML置标;

⑦ 检索要求:提出元素对应属性字段的具体检索要求;

⑧ 系统需要:元数据需要作为期刊论文服务系统需求的一个部分;

⑨ 未来扩展:预留未来发展的可能性,即可扩展性。

3 期刊论文元素标引规范研究

3.1 标引规范的内容结构分析

期刊论文标引数据的基本结构分为三个部分,即核心元素、专用元素和特定元素^[3]。通过这样的分类方法,可以很明确地知道哪些数据是数字资源对象通用的,哪些数据是期刊论文专用的,哪些数据是特定领域的期刊论文专用的,便于分类标引和检索。

(1) 核心元素

核心元素是指在各类出版物资源对象的描述中都通用的元素。核心元素的应用对资源类型没有限制,可以在会议论文、期刊论文等出版物多种资源中使用,具有简单、通俗易懂、扩展灵活、通用性强等特点。比如“题名”、“作者”这类元素。

(2) 专用元素

专用元素是指在期刊论文资源对象中通用的元素。期刊论文是期刊的析出文献,是期刊中相对独立的一部分。对期刊论文的描述不仅包括描述对象本身,还包括期刊论文之间的关系描述,同时还要提供对其母体文献期刊的简单描述。比如“期刊”这类元素,不仅包括期刊本身的信息,还包括期刊论文的起止页码。

(3) 特定元素

特定元素以某种特定的数字资源对象的属性、特点为基础制定,仅适用于这类对象,可不用于交换,应用时仅要求该种对象遵守其定义。在本规范研究中,特定元素是针对某些特定种类期刊的元素,比如“医药学专有名词”只针对医药类期刊论文。

数字化期刊元素标引规范,将其层次关系分为四层,分别为元素、子元素、修饰词、扩展修饰词。由于期刊论文信息庞杂,故需要对有些元素进行详细的属性扩展,子元素对元素进行限定说明,修饰词对子元素进行限定说明,扩展修饰词对修饰词进行限定说明。对元素的语义进行逐级修饰,提高元素的专指性和精确性。标引粒度更细,能够实现对庞杂的期刊论文信息的有效控制,使海量信息有序化,便于检索利用。每个数据会被归属为核心元素、专用元素、特定元素中的一个。

3.2 标引规范的元素构成分析

期刊论文的元素标引主要包括核心元数据标引、文档结构标引和深度内容标引三个方面,如图1所示。

元素标引阶段是指在期刊论文的写作、编辑、排版、加工等不同的环节进行元素的标引,可以分为作者阶段、编辑阶段、排版阶段和加工阶段。在作者阶段,由作者对其写作论文的初始元素进行标引,包括论文题名、作者、关键词、参考文献、文档结构、知识点等,尤其是图、表、公式与论文内容之间的映射和关联关系等,只有作者最清楚,所以需要在作者阶段标引。在编辑阶段,由编辑对作者投稿的期刊论文进行审核,对其修改和添加的元素进行标注,包括期刊的基本信息和出版信息、期刊论文的DOI等。在排版阶段,对期刊论文的相关排版元素进行标注,如论文在期刊中的起始页码等。在加工阶段,对数字期刊论文的全文加工过程中对其关键元素进行标注,包括机标关键词、机标分类号等。

3.2.1 期刊论文核心元数据标引

期刊论文的核心元数据包括期刊及其论文的著录信息、关键词、分类法和参考文献等常规项的标引元数据规范,与以前的期刊论文标引规则相比,要求标准化程度更强,标引粒度更细。借鉴国内外元数据应用项目的实施方案,数字期

刊论文应包括以下元素,每个元素都有自己的元数据^[4]:

(1) 作者元素:期刊论文的作者,包括作者标识、论文标识、作者名称、作者单位、作者简介等元数据。

(2) 期刊元素:与论文及其发表期刊相关的元数据,包括期刊标识、论文标识、论文名称、期刊名称、年卷期、论文摘要、关键词、分类号、文献类型等。

(3) 基金元素:与基金项目相关的元数据,包括基金标识、论文标识、基金项目名称、基金项目编号等。

(4) 参考文献元素:与论文引用和被引相关的元数据,包括论文标识、引文论文标识、引文作者、引文标题、引文期刊标识、年卷期等,参照国家标准GB/T 7714—2005文后参考文献著录规则^[5]。

3.2.2 期刊论文文档结构标引

期刊论文文档结构包括各级标题、段落、脚注尾注、页眉页脚、图、表格和公式等文档格式的标引和识别规范,主要包括以下元素:

(1) 标题元素:论文各级标题的样式,包括各级标题的字体、缩进、行距、段前段后等。

(2) 段落元素:论文段落的格式,包括正文各段的段落缩进、行距、段前段后等。

(3) 脚注元素:论文脚注和尾注。

(4) 页眉页脚元素。

(5) 图、表格、公式的识别和元素标引。

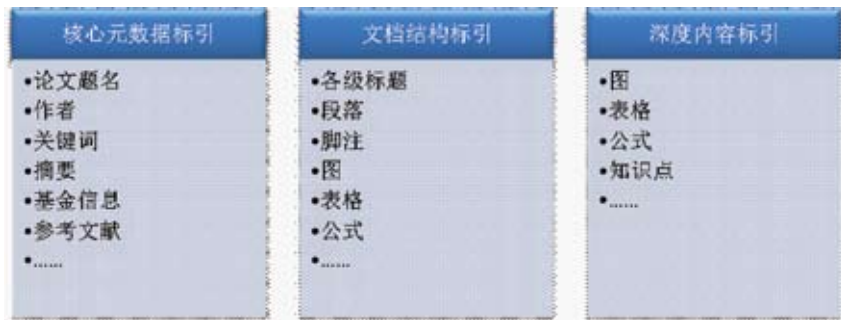


图1 期刊论文元素标引

3.2.3 期刊论文深度内容标引

目前对期刊论文的标引,主要集中在著录数据和主题标引的研究上,忽略了图、表、知识点、专有名词等深度内容信息,这样就导致了期刊论文中很多信息的遗失。

(1) 知识点标引

在有些信息活动中,除了期刊论文所含信息之外,还涉及到论文中的知识和事实。因此仅对论文进行信息标引是不够的,知识点标引就是用于对科技文献中的知识和事实的标引,我们称之为知识点标引。

知识点标引是实现跨领域知识集成与知识发现的基础,主要包括以下描述型和过程型两大类型的知识元,描述型知识元主要包括数值、术语等,过程型知识元主要包括方法、步骤、原理等^[6,7]。

(2) 专有名词标引

专有名词指表达某一单一事物对象的名称。常用的专有名词包括人名、地名、机构名、团体名称、会议名称、作品名称及各种专业术语等^[8],数量大,范围广。不少专有名词具有巨大的标引和检索价值。

(3) 图标引

一篇论文有可能包含一至多张图,目前大多数文献标引都没有对图进行标引,但这方面的研究和应用已引起很多资源单位的关注,有些已开始尝试对文献中的图进行标引,但也仅从图像名称、图像尺寸、压缩类型、作者、年代等方面标引图,对图元素所蕴含的内容和知识没有作详细分析。因此,本规范所研究的图标引元数据,除了能够描述对图的基本信息,还要建

立起图中的元素与论文内容之间、图与原始数据之间的映射和关联关系。根据目前期刊论文的实际情况,我们将图分为以下三种类型:

1) 数据生成图

定义:由原始数据生成的图。

举例:图表、谱图、聚类图。

2) 影像图

定义:由影像编辑软件或经由扫描仪、影像光盘、数字相机等设备而产生的图。

举例:数码照片、卫星影像、扫描图片。

3) 绘制图

定义:采用制图软件绘制的图

举例:流程图(Visio)、建筑图(CAD)

(4) 公式标引

基于目前的技术,能够采用数学置标语言(Mathematical Markup Language, MathML)来描述数学表达式,但却不能对公式及其符号所表达的含义进行描述。因此,需

要制定一套不仅能描述公式形式,而且能准确、全面揭示公式含义的元数据规范。

(5) 表格标引

表格和图像、公式一样,是科技文献中的常用资源,通常用以比较详细资料和数据的不同。对表格的标引同样也不能仅限于基本信息描述,应对表格及其元素(资料或数据)的含义进行描述,即建立起表格与论文内容之间、表格与图像之间(特别是数据表)的映射和关联关系。

3.3 元素标识与元数据的互操作框架

数字化期刊元素标引规范基于文献元素标识与元数据库之间的互操作框架,规定系统在用户进行元素标引的同时,与元数据库建立映射关系,并对元数据库的相应记录进行添加和更新,如图2所示。

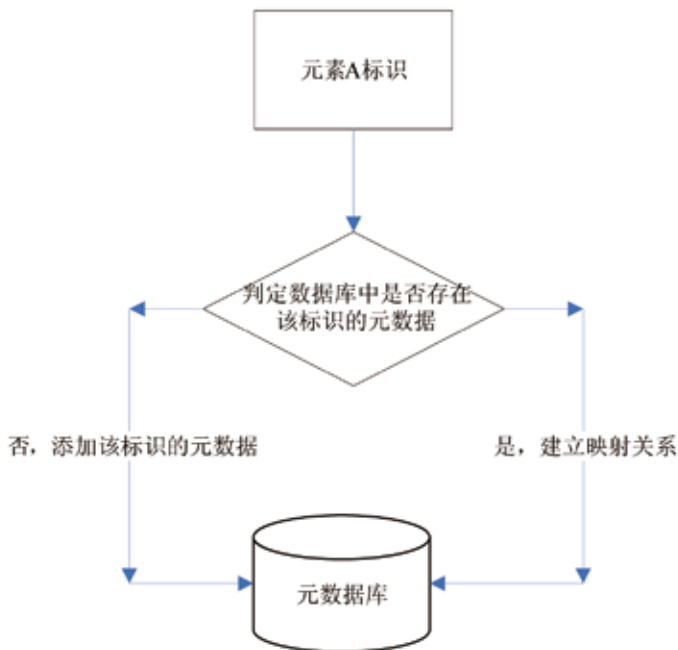


图2 互操作框架

4 结论

论文元素及其元数据对于用户使用期刊论文具有巨大的应用价值,但是目前期刊论文各类信息的标引程度不高,使得大量信息流失而不能得到有效应用,期刊论文元素标引规范通过对期刊论文,尤其是数字期刊论文的重要元素进行描

述,对期刊论文本身及在写作、编辑、排版、加工、利用过程中产生的二次信息加以规范化、开放性的揭示,有助于用户或第三方系统方便地、无障碍地共享期刊论文中的有用信息,有助于各个论文服务系统通过元素的元数据,有效地交流论文基础信息及深层内容信息,促进不同系统之间的互操作

性。在下一步研究工作中,将对标引规范进行完善和应用实践,在标引规范制定的同时,加强标引和识别具体技术的研究,结合标引方法研究开发转换工具,将在写作编辑阶段完成的内容深度标引转换为XML规范文档,满足未来灵活使用的实际需求。

参考文献

- [1] 我国数字图书馆标准规范建设项目组. 专门数字对象描述元数据规范设计指南[R],2004.
- [2] 刘炜. 数字资源组织的元数据方法[C]//中国图书馆学会2005年年会论文集,2005.
- [3] 我国数字图书馆标准规范建设项目组. 期刊论文描述元数据规范[R],2004.
- [4] 钱建立,吴广茂,张薇. 中文期刊论文元数据的选取和实现[J]. 情报杂志,2008,27(10).
- [5] GB/T 7714—2005文后参考文献著录规则[S].
- [6] 温有奎,温浩,徐端颐,等. 基于知识元的文本知识标引[J]. 情报学报,2006,25(3).
- [7] 原小玲. 基于知识元知识标引[J]. 图书馆学研究,2007,(6).
- [8] 李楠. CNMARC环境下专有名词及专指主题词的标引特点[J]. 中华医学图书情报杂志,2010,19(3).

作者简介

吴琳(1981-),主要从事数字出版领域的研究工作,发表论文数篇。E-mail: wulin@wanfangdata.com.cn

The Element Indexing Standard Research of Journal Articles

Wu Lin / Institute of Scientific and Technical Information of China, Beijing, 100038
Cheng Yuhua / WANFANGDATA CO., Ltd, Beijing, 100038

Abstract: Based on journal articles characters and application request, this paper analyzed journal article element indexing significance and designed process for journal article element indexing establishment standard. Then the paper deeply analyzed and researched indexing standard content structure, element composing and inter-operation frame. According to this, the paper proposed journal article indexing elements on core metadata, document structure, deep content, and pointed out the directions for next stage researching work.

Keywords: Journal article, Element indexing, Indexing standard, Metadata indexing

(收稿日期: 2010-11-17)