NSTL统一文献元数据标准的设计与思考*

张建勇，于倩倩，黄永文，董智鹏
（中国科学院文献情报中心，北京 100190）

摘要：分析了NSTL统一文献元数据标准建设的必要性，介绍了统一文献元数据的设计目的是为保证NSTL发展战略目标的实现。元数据的适用对象涵盖NSTL所有科技资源。元数据的设计原则包括前瞻性原则、协同化原则、最小粒度原则、模块化原则和兼容国际标准原则。提出元数据设计思路，并详细介绍了其中的功能需求分析，构建了领域模型。本研究在元素和属性的选取方面主要参考JATS标准。

关键词：NSTL；元数据；JATS；设计
中图分类号：G250.7
DOI：10.3772/j.issn.1673-2286.2016.2.005

1 引言

当前，数字出版已经成为科技文献资源的主要出版形态，描述科技文献的元数据规范日渐增多，有些是国家标准，有些是公司内部标准。例如NISO JATS Version 1.1[1]作为美国国家标准，得到了广泛应用和认可[2]；科技平台资源核心元数据[3]于2014年成为中国国家标准，为国家科技基础条件平台门户提供统一的元数据；Web of Science[4]、Scopus[5]作为具有较大影响力的数据库，其元数据规范已成功应用于数字化生产过程；Dryad元数据规范[6]被称为科学数据仓储元数据的最佳实践；DC元数据[7]具有较强通用性，但相对来说数据元素简单。纵观现有元数据规范，虽各有特色，却也有很多相同之处。如通过一套Schema描述多种类型文献，数据项丰富、多用属性进行描述，具有多种唯一标识符等。这也较好地印证了大数据时代，资源组织颗粒度细化、资源灵活挖掘与整合日益重要的特点。

国家科技图书文献中心（以下简称NSTL）经过多年的发展，已经形成从采购、加工、发布到服务的数字化业务流程，各个子系统相互协同、相互依赖，共同为用户提供服务。但在发展过程中，各个层面和系统都制订了自己的元数据方案[8-10]，导致NSTL各层面系统使用的元数据规范不尽相同，难以实现资源的深度挖掘，并限制了系统的可持续发展。NSTL近两年通过赠与、呈缴、购买等方式获取了国内外出版商和相关信息机构的元数据并进行应用[11]，但这些来自出版商、数据库商和服务商的元数据遵循的标准各有差异，对资源的共享和利用造成障碍。只有将NSTL各层面系统产生和转换自第三方来源的数据进行整合后，才能形成可进行大数据存储管理、分析挖掘的数据，建立NSTL统一文献元数据标准将有利于大数据基础设施的建设。

2 设计目的与对象

NSTL“十三五”发展规划提出，要全面构建国家科技文献信息大数据管理与服务体系。在大数据时代，元数据的重要性毋庸置疑，数据能被拆分、重组、分析和挖掘，都需要元数据的参与。建设NSTL统一文献元数据标准，能够支持多种数据的统一描述，形成一致的数据描述体系，推进科技文献信息深度组织和揭示，将为NSTL数据集成融合、数据分析和数据挖掘，以及不同应用服务系统间的互操作打下数据基础，从而为科学决策和知识服务提供支撑。

NSTL统一文献元数据标准的总体设计目标是为

* 本研究得到NSTL项目“建立和发展NSTL元数据标准规范体系”（编号：2015XM04）资助。

NSTL建成国际一流的科技文献信息发现和保障体系，实现从信息服务向知识服务的转型，提供数据标准的基础保障，保证NSTL发展战略目标的实现。具体目标为支持NSTL文献发现系统的建设，支持数据挖掘、分析评价功能的实现，支持系统间数据交互的可靠性，保证各个层面系统数据重用和利用的标准化，降低系统间数据传递损失，增强系统间的协同能力。

NSTL统一文献元数据标准的设计对象涵盖所有的NSTL购买、交换、赠与等方式获取的科技类资源，包括图书、期刊、会议录、期刊论文、会议论文、学位论文、科技丛书、工具书、文集汇编、科技报告、开放课程、开放课件等。可统一描述文献的印刷版本、数字版本，统一描述文献对象各个层次的信息，满足NSTL数字业务流程中文献数据采集、管理和服务的需求。

3 设计原则

随着数字信息资源的普及和相关技术工具的成熟，数字信息本身的解析颗粒化，以及关联和重组的特性开始全面影响信息资源的组织和利用，元数据描述也呈现出细粒化、结构化、语义化和关联化等发展趋势。NSTL统一文献元数据标准的设计必须与时俱进，既要考虑新的形势，又要考虑可能的潜在需求，设计原则如下。

3.1 前瞻性原则

NSTL“十三五”发展战略明确了从文献传递服务为主向资源发现服务、分析评价服务转型，从文献保障为主向知识服务基础支撑保障转型的发展方向。统一元数据标准规范的设计应充分考虑NSTL未来5年或更长时间的发展需求，数据标准规范不仅支持资源的发现，也支持基于数据的分析评价和知识服务的要求。在元数据的设计上不仅考虑揭示文献的基本信息，也考虑揭示全文层面的图表和公式等信息，同时也预留了全文描述字段内容。在设计上充分考虑服务的扩展和深入发展的需要。

3.2 协同化原则

统一文献元数据标准规范的设计目的是满足NSTL数字业务系统中各个子系统应用的需要，各个系统可采用同一个标准描述文献对象，各个系统可以基于自己的管理需要描述文献对象的不同深度的内容，但遵循同样的数据标准，为后续数据的复用和深入加工建立良好的基础。例如对一篇期刊论文的描述，数据格式应是统一的，编目系统的描述和数据加工系统的描述最后应统一成一个数据标准描述，最后形成的数据满足资源发现和分析评价的需要。统一元数据标准规范的设计充分考虑各个子系统的特点，在数据模型和数据描述上支持各个子系统协同管理的需要，各个子系统通过协同达到最大的数据管理效益。

3.3 最小粒度原则

数据描述的粒度越小，数据描述越精确，可供分析评价的点就越丰富。统一文献元数据标准规范确定的数据描述粒度尽可能细致到原子层面，按最小粒度设计元素或属性，以支持下一步分析评价和知识服务的需要。例如机构字段，可细分为一级机构名称、二级机构名称、所在国家、城市、地址等，这样描述为下一步精确定位机构和统计分析机构的产出建立基础。在统一文献元数据的设计中，最小粒度原则贯穿在各个层面，尽可能细致地描述文献对象的各个层面信息，为下一步数据的分析评价打下基础。

3.4 模块化原则

模块化是现代元数据设计最重要的特征，根据实体关系方法分析抽象出资源对象的实体关系模型，对资源的描述就是对模型中不同实体进行描述再组合而成。领域模型中具有共同特点的实体对象可复用描述不同层面的数据对象，例如机构实体，实际上可以是研究者所在机构，也可以是出版机构、资助机构和学位授予机构，机构的元素构成是一致的，成为一个公用的实体模块在描述中使用，也为下一步数据管理规范打下基础。

3.5 兼容国际标准原则

国外部分大型出版机构已经建立相关的文档结构规范，并且具有完整的描述体系结构。例如NLM制订的JATS标签集经过多次修改，已经成为美国国家标准并发布了最新版本ANSI/NISO Z39.96-2015[1]。统一文献元数据标准应与国际上主流的相关国际标准兼容，

以便融入国际数据大环境。

4 NSTL统一文献元数据标准设计思路

根据NSTL文献元数据制订指南[12]确定的设计元数据的通用技术框架和其定义的流程方法，并面向应用、面向服务对NSTL统一文献元数据进行设计与建设。基本流程包括功能需求分析、领域模型分析、设计元数据记录、编制使用指南、元数据形式化描述。在这个流程中更多强调需求分析和领域模型分析，元数据记录的设计基于元素和属性的方式构建，强调元素定义的一致性和包容性，可描述多样化、多层次的资源。

4.1 功能需求分析

功能需求分析主要是描述设计元数据需要满足的具体应用需求。统一文献元数据标准支持NSTL文献发现系统的建设，支持数据挖掘、分析评价功能的实现。NSTL文献数据库包括期刊论文、会议论文、学位论文、文集汇编、科技报告等。期刊论文、会议论文、文集汇编都是集结出版的文献，学位论文和科技报告则通常是单篇或者成册出版。NSTL文献数据库元数据从功能上应支持以下功能。

4.1.1 文献检索和选择

即满足用户根据特定条件检索、选择文献并对文献进行排序的需求。包括：①按类型如图书、期刊、科技报告等检索选择文献；②根据文献主题和内容如题名、关键词、主题词、摘要等检索选择文献；③根据文献特征和特定条件如作者、作者机构、ISBN、ISSN等检索和选择文献；④根据文献引用频次选择文献。

4.1.2 文献识别

即对各类文献内容特征和外部特征进行描述。包括：①根据文献特征如文献的唯一标识符识别；②识别文献作者及其所在机构，如通过orcid、researcherID识别文献作者，通过机构唯一标识符识别机构等；③通过全球通用的DOI识别文献；④通过NSTL本地通用的Local ID识别文献；⑤识别全文的版本和载体形式如印本、电子版本等。

4.1.3 全文获取

即满足用户对印本和电子版本全文的获取需求。包括：①支持在NSTL九家成员馆范围内的全文获取；②支持对各种载体和版本全文的获取，提供能够链接到全文的多种选择；③支持对开放获取全文文献的获取。

4.1.4 文献分析评价

即从不同方面对文献进行分析评价，满足用户对科研产出分析挖掘的需要。包括：①支持引文关系的描述和计量名称识别；②支持对人名、机构、资助者和项目的产出分析评价的需求；③支持面向学科的文献分析评价。

4.1.5 使用授权

即针对来自不同机构的不同用户，文献可获取方式和获取范围的授权有所不同。包括：①文献的印本馆藏信息和网络版本获取授权方式；②来自作者、出版社和其他各个方面的开放获取资源的授权信息。

4.1.6 内部数据管理

即对内部数据采集、描述、保存方面的管理，能够及时掌握数据的动态，如遇特殊情况，能够及时修复数据。包括：①数据产生、更新、删除等时间责任人记录；②描述数据状态和数据层次；③支持数据审计。

4.2 领域模型分析

通过分析元素集合及其相互之间的关系，构建领域模型。根据实体分析方法，对期刊、图书、会议录、科技丛书、期刊论文、会议论文、学位论文、文集汇编、科技报告等各类资源进行研究和分析。可以发现，NSTL统一文献元数据可以分为12个元素集，包括来源元素集、论文元素集、全文元素集、引文元素集、图表元素集、附加资料元素集、Agent元素集、主题元素集、基金元素集、会议元素集、获取管理元素集和操作信息元素集。

其中，来源元素集主要是描述期刊、图书、会议录

等来源信息；Agent元素集包括贡献者和机构信息，贡献者可以是作者、编辑者和指导人员等，机构可以是作者所属机构、著作的出版机构、会议的举办机构和基金项目的资助机构；获取管理元素集主要描述获取方式和使用授权信息；操作信息元素集描述数据的更新、处理状态等。NSTL统一文献元数据的领域模型如图1所示。

NSTL统一文献元数据的领域模型中元素集之间的关系可以概括为5种关系：

（1）文献内部元素集之间的关系。一个来源可以包含一篇或多篇论文，一篇论文可以有一个或多个全文，一篇论文可以有一个或多个引文，一个全文可以有一个或多个图表、有一个或多个附加资料。

（2）文献与其他元素集之间的关系。一篇文献可以有一个或多个贡献者，一个贡献者可以属于一个或多个机构，一篇文献可以由一个或多个机构出版，一篇文献可以由一个或多个基金资助，一篇文献可以发表在一个或多个会议上，一篇文献可以有一个或多个主题，一个会议可以由一个或多个机构负责举办，一个基金项目可以由一个或多个机构资助，一篇文献可以有一个或多个获取管理和操作信息。

（3）元素集与规范记录之间的关系。来源、主题、Agent、基金、会议可分别对应一个规范记录。

（4）来源与来源之间的沿革关系，主要包括继承、部分继承、替代、部分替代、吸收、部分吸收、分自等关系。

（5）文献与文献之间的关系，主要包括引用关系、相似关系等。

4.3 设计元数据记录

设计元数据记录首先需要确定元素和属性，NSTL统一文献元数据元素和属性的选取、定义主要参考NISO JATS 1.1[1]。一方面因为JATS作为美国国家标准，应用广泛。例如出版商、知识库、图书馆、软件开发商、学术机构、期刊等身份不同的机构支持JATS的使用和推广[2]。NSTL接收的第三方来源元数据包括CUP[13]、OUP[14]、De Gruyter[15]等也采用了JATS标准，参考JATS便于NSTL与第三方来源元数据的交互。另一方面因为JATS可以描述到全文，为下一步扩展留下了足够的空间。

在JATS中，元素通常为名词，代表了文献的一部分，例如题名、摘要、作者等。属性更进一步地对元素进行描述，例如使用 xml:lang属性表达语种信息，使用article-type属性表达文献类型信息等。每个属性都会有属性名和属性值，属性可以对表达相同内容的元素进行归并。对于JATS中与NSTL需求相同的元素和属性，进行复用，并保持语义的一致性，对于NSTL有实际应用需求而JATS未定义的元素或属性，进行扩展，扩展的元素或属性不与JATS发生冲突，在元数据的描述结构

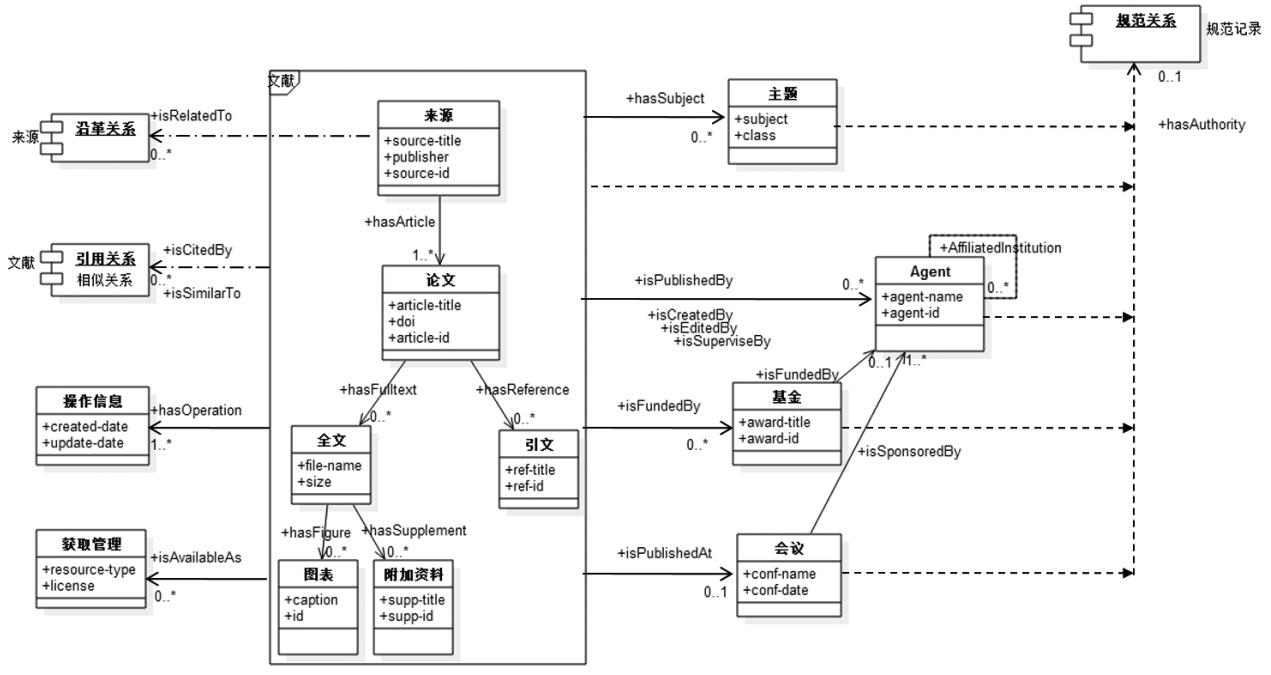

图1 NSTL文献元数据领域模型

上尽量与JATS保持一致。

在确定元素和属性后，对元数据记录进行设计和描述。设计元数据记录需要考虑的问题包括元素的出现频次、元素取值、编码体系、元素出现的顺序、元素间的交叉引用关系等相关的技术细节约束。根据NSTL文献元数据制订指南要求，分别从12个方面对元素进行定义（见表1），从5个方面对属性进行定义（见表2）。

4.4 编制使用指南

使用指南提供元数据的著录规则，解释原因并指导人们创建元数据。理想状态下，使用指南解释每个元素，预测在元数据创建过程中产生的问题并作出指导。使用指南中包含与元数据记录结构中相同的一些信息，但相对来说更便于人理解。使用指南中可能包含的规则如对作者进行著录时，若成果中包含有多个作者，则选择前三个进行著录；关键词的著录参照某种规范等。

4.5 元数据形式化描述

形式化描述是以计算机可读方式描述规范，通常使用计算机语言如XML语言、RDF语言等对元数据进行形式化描述。在NSTL统一文献元数据的设计中，考虑到现有元数据规范通常采用XML语言作为编码和数据交换语言，本标准也采用XML语言实现元数据的形式化描述。XML语言包含了一组定义语义标记的规则，可以定义特定领域内标记语言的语法结构。

5 结语

在大数据和新型数字信息环境下，如何应对海量

表1 元素定义表

序号	名称	英文名称	说明	约束
1	名称	Name	为方便计算机处理而定义的元素标记，名称通常使用英文	必备
2	标签	Label	通过适合人们阅读的词汇描述元素，标签一律使用中文	必备
3	URI	URI	元素的唯一标识	必备
4	定义	Definition	对元素含义的解释性说明文字	必备
5	出现频次	Occurrence	0..1，1，0..*，1..*	必备
6	注释	Remarks	对元素的附加性说明	有则必备
7	内容表示	Content model	dtd或xsd表示	必备
8	扩展的内容表示	Expanded Content model	扩展内容的dtd或xsd表示	有则必备
9	描述	Description	元素间的结构关系	有则必备
10	相关元素	Related-element	与该元素相关的元素	有则必备
11	属性	Attribute	元素中应用到的属性	有则必备
12	示例	Example	xml样例说明	有则必备

表2 属性定义表

序号	名称	英文名称	说明	约束
1	名称	Name	为方便计算机处理而定义的属性标记，名称通常使用英文	必备
2	标签	Label	通过适合人们阅读的词汇描述属性，标签一律使用中文	必备
3	定义	Definition	对属性含义的解释性说明文字	必备
4	属性值	Value	属性的取值内容，包含编码体系	必备
5	使用限制	Constraints	属性使用的限制	有则必备
6	示例	Example	xml样例说明	有则必备

数据的产生和分析挖掘成为挑战。NSTL已有的数据和下一步要建设的数据来源于多个系统和渠道，数据描述标准和格式多样，由此带来的复杂问题对后期的数据分析管理极为不利。本文对NSTL统一文献元数据标准的设计目的、对象和原则进行了介绍，并提出设计思路，主要包括功能需求分析、领域模型分析、设计元数据记录、编制使用指南和元数据形式化描述，希望能够为相关信息系统的元数据建设提供参考和借鉴。

参考文献

[1] NISO JATS Version 1.1 (ANSI/NISO Z39.96-2015) [EB/OL]. [2015-12-22]. http://jats.nlm.nih.gov/archiving/tag-library/1.1/index.html.

[2] 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.

[3] 科技平台资源核心元数据[EB/OL]. [2015-12-22]. http://www.most.gov.cn/ztzl/kjzykfgx/kjzykjptbz/kjzybz/201407/t20140718_114487.htm.

[4] ISI Web of Science-Science Citation Index Expanded [EB/OL]. [2015-12-22]. http://www.webofknowledge.com/WOS.

[5] Scopus [EB/OL]. [2015-11-20]. http://www.scopus.com/.

[6] Greenberg J, et al. A Metadata Best Practice for a Scientific Data Repository [J]. Journal of Library Metadata, 2009, 9(3-4): 194-212.

[7] DC metadata [EB/OL]. [2015-12-22]. http://dublincore.org/.

[8] 张建勇,曾燕.文献数据库数据加工规范[M].北京:知识产权出版社,2009.

[9] 吴思竹,胡铁军,梁芳,等.NSTL联合目录系统元数据的数据逻辑结构设计[J].图书馆杂志,2014(1):31-35.

[10] 翟爽,赵艳,王昉.NSTL开放课件元数据规范及一体化建设研究[J].数字图书馆论坛,2015(8):22-27.

[11] 于倩倩,张建勇.NSTL集成利用第三方来源元数据的实践与探索[J].现代图书情报技术,2016(1).

[12] 张建勇,于倩倩,黄永文,等.NSTL文献元数据制订指南(内部资料)[R].2015.

[13] Cambridge Journals [EB/OL]. [2015-12-22]. http://www.journals.cambridge.org.

[14] Oxford University Press [EB/OL]. [2015-12-22]. http://www.oxfordjournals.org/en/.

[15] De Gruyter [EB/OL]. [2015-12-22]. http://www.degruyter.com/.

作者简介

张建勇，男，1965年生，中国科学院文献情报中心研究馆员，研究方向：数据库建设和数据管理。
于倩倩，女，1986年生，中国科学院文献情报中心助理馆员，研究方向：数据管理和组织，E-mail：yuqianqian@mail.las.ac.cn。
黄永文，女，1975年生，中国科学院文献情报中心副研究馆员，研究方向：数据管理和组织。
董智鹏，男，1985年生，中国科学院文献情报中心馆员，研究方向：文献数据管理。

Metadata Standard Design of NSTL Unified Literature

ZHANG JianYong, YU QianQian, HUANG YongWen, DONG ZhiPeng
(Library of Chinese Academy of Sciences, Beijing 100190, China)

Abstract: This paper analyzes the design necessity of metadata standard for NSTL unified literature. The objective of the metadata standard design is to ensure the implementation of NSTL development strategy. The designed metadata standard can be applied to all NSTL S&T resources. The proposed design principles include prospective principle, collaborative principle, minimal granularity principle, modular principle, and compatible principle. This paper comprehensively demonstrates the design solution of metadata standard, introduces the functional requirement of the standard, and constructs the domain model. Additionally, selection of metadata element and attribute is based on the JATS.

Keywords: NSTL; Metadata; JATS; Design

（收稿日期：2016-01-22）