数据溯源模型与DC元数据的映射研究

林芳芳，赵辉
（中国科学技术信息研究所，北京 100038）

摘要：PROV是关于数据溯源的一系列规范，与DC元数据存在密切的联系。在数字图书馆领域，数据溯源成为一种趋势，如何利用DC元数据建立数据溯源体系就成为研究重点之一。本文从DC元数据和PROV概念入手，比较分析其关系，探讨两者之间的映射，得出若要满足数据溯源要求，DC元数据需增加描述活动、关系、代理相关字段的结论。

关键词：数据溯源；PROV；DC元数据；映射；数字图书馆
中图分类号：TP391
DOI：10.3772/j.issn.1673-2286.2016.3.002

在大数据背景下，数据成为重要的资产。人们希望能够像实物资产一样，在数据使用过程中，一旦出现质量问题，可以根据溯源信息，找到数据产生和生产环节中可能出现问题的地方，提高数据生产和使用的效率和效益。因此，“数据溯源”的概念应运而生。“数据溯源”也叫数据起源、数据族系，是对数据的追本溯源，不仅强调数据溯源追踪技术，实现对历史数据的重现，更强调从原始数据到数据产品衍生的过程。PROV作为2013年W3C出的数据溯源标准，提供以数据溯源模型（PROV-DM）文档为核心的12个系列文档（包括4个推荐标准），实现对数据的溯源及规范化表达。其实现原理是通过捕捉溯源所需的相关数据，利用数据溯源模型（PROV-DM）和数据溯源本体（PROV-O）分别进行组织及表达。数字图书馆中包含大量的数据资源，是科研工作者在创新活动中要使用的不可或缺的资源。数字图书馆引入数据溯源标准，进一步加强对数字资源的管理，将更有利于数据资源的开发和利用。元数据对数字资源进行描述、组织、管理，在数字图书馆建设和管理中起重要作用。DC元数据因其在数字资源描述上的简易性、通用性、可扩展性等特点而被大部分数字图书馆采用。但现有的DC元数据是否满足数据溯源的要求，亟需研究和确认。本文将数据溯源标准PROV与DC元数据标准进行比较，考察两者的联系和区别，为数据溯源标准在数字图书馆的应用提供支撑。

1 相关研究述评

数据溯源是一个新兴的研究领域，国外针对数据溯源的研究主要集中在方法、模型以及应用三个方面。方法研究上，提出的常用方法有注释法[1]、反向查询法[2]。模型研究上，提出的通用模型包括OPM模型[3]、Provenir模型[4]以及最新的PROV-DM模型[5]等，此外国内学者在各类模型基础上提出一些改进，如OPM安全扩展模型[6]、DNA双螺旋模型[7]。应用上，早期主要集中在生物、天文、地球科学、地理信息系统等专业领域，后来逐渐扩展到计算机等通用技术领域。目前，部分学者研究和应用W3C发布的PROV系列文档，致力于补充或完善该文档定义的PROV-DM模型使其面向特定领域使用，或将该文档描述的数据溯源词汇与相关领域词汇进行映射，其中包括PROV与音乐本体间的映射[8]。

国内对于数据溯源的研究主要集中在计算机和国防领域，图情领域也有相关研究。图情领域针对数据溯源展开的研究有邓仲华等对面向数据发布的科学工作流数据溯源方法进行研究[9]；李文燕等分析比较了常用的溯源模型OPM、Provenir和CRMdig，对溯源标准PROV进行研究和分析[10]；倪静等则详细分析了PROV标准中的PROV数据溯源模型及其Web应用，并对Web应用中溯源信息定位和查询机制进行相关研究[11-12]；吴振新等通过分析长期保存领域相关标准中

对溯源的要求和描述来研究溯源技术在长期保存中的应用，提出长期保存溯源管理框架[13]。

综上所述，数据溯源已经成为国内外研究者关注的一个领域。利用数据溯源思想和模型对数字图书馆中的数字对象进行进一步管理也成为数字图书馆的研究热点。研究数据溯源PROV与DC间的映射关系，有助于开发人员从大量DC数据中提取PROV数据，使DC术语中包含的溯源信息更加明确，提高DC和PROV的互操作性，也有助于数字图书馆资源更加适应大数据的应用环境。

2 DC和PROV的基本概念

数字图书馆中常用的DC元数据标准包含15个核心元素和限定词。PROV从数据溯源的需求出发，提出数据溯源的概念模型和使用规范，其具体使用时必然要与DC元数据建立映射，并补充已有元数据项的缺失，而后才能建立起数字图书馆数据溯源的技术体系。

2.1 DC（都柏林核心元数据）

DC是 Dublin Core的简称，是1995年3月由OCLC和NCSA在美国俄亥俄州都柏林召开的第一次元数据研讨会上提出的概念，用以描述资源对象。DC是在网络资源迅速增长下出现的一种描述性元数据，包括15个核心元素，较全面地涵盖了数字资源的主要特征，能够很好地描述和揭示数字资源。利用DC元数据能对信息资源进行描述、定位、评估、选择，是描述、管理和检索数字资源的有效组织方式[14]。在应用上，已经从单纯的数字资源描述，扩展到数据资源的管理活动描述、技术管理描述等多个方面。经过20年的发展，DC元数据已成为数字图书馆中信息资源组织和管理的重要工具。

2.2 PROV

PROV是2013年W3C面向用户、开发人员和高级开发人员发布的关于溯源的标准[15]。目前，W3C PROV工作小组共发布12个文档，其中有4个作为推荐性标准（见表1），其余8个作为工作草案。PROV-DM作为该系列文档的核心，定义了一个通用的、与领域无关的溯源模型，同时发布了3个接口性规范：PROV-N设计了一种人可读的记录溯源的符号来表达溯源模型；PROV-O即PROV本体是用OWL2网络本体语言对PROV-DM进行编码，实现PROV-DM提出的概念模型；PROV-CONSTRAINTS则指出确定有效的溯源实例必须满足的约束条件。为便于数据溯源信息的使用，还提供了PROV-AQ协议来定位和访问溯源信息。

表1 PROV推荐标准

序号	推荐标准	说明
1	PROV-DM	溯源信息数据模型，该序列化文档包括PROV-O、PROV-XML（工作草案）和PROV-N
2	PROV-N	为PROV-DM定义了人可读的记录溯源信息的符号
3	PROV-O	用OWL2描述从PROV-DM到RDF的映射
4	PROV-CONSTRAINTS	使用PROV-DM的限制条件

其中，PROV-DM是概念数据模型，是W3C溯源系列规范（PROV）基础。PROV-DM核心结构由“三类七关系”构成，包括Activity、Entity、Agent三类和Used、WasAssociatedWith、WasAttributeTo、WasDerivedFrom、WasInformatedBy、WasGeneratedBy、ActedOnBehalfOf七个关系。该模型描述了Activity（活动）使用Entity（实体）或产生新的实体，并通过Agent（代理）来控制活动的过程，如图1所示。PROV-DM核心结构关注于溯源管理对相关事项的描述，并用有向图的形式表示“谁（Agent）对实体（Entity）做了些什么（Activity）”的信息。

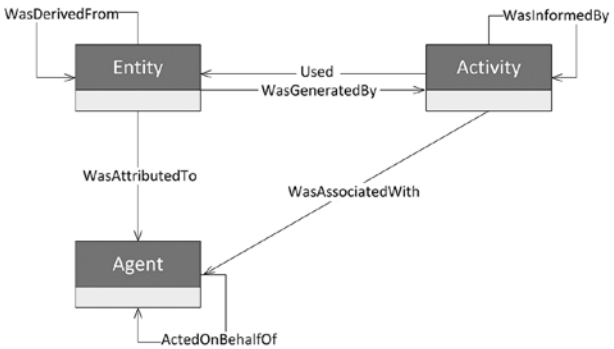

图1 PROV推荐标准

2.3 PROV与DC元数据的关系

都柏林核心元数据计划（The Dublin Core Metadata

Initiative，DCMI）提供核心元数据元素集（通常称为DC）用于简洁和通用的资源描述，是物理对象数字化和数字资源管理的基础标准，这是现实世界中的事物纳入数字环境下进行管理的第一步。图书馆领域已经根据自身对数据对象管理的需要，形成了一套以DC元数据为基础的元数据。在这些元数据项中，一部分可以被溯源模型直接采用，如dataAccepted、dataCopyrighted、issued等。但利用PROV提供的数据溯源模型，从数据溯源的角度来看，数字图书馆中已有的DC元数据在描述实体、活动、代理等相关元素及其关系时，还有不完备之处。因此，建立DC与PROV之间的映射，从DC元数据中提取实现数据溯源所需信息，是数据溯源模型在数字图书馆应用的第一步，也是关键的一步。

3 PROV与DC之间的映射

DCMI[16]2008年发布了“DCMI Metadata Terms”，开始启用“术语（Term）”和“命名空间（Namesapce）”来扩展原来的元素和修饰词，其中将原来的元素、修饰词、编码方案统称“术语”。本节首先梳理出DC术语中相关的溯源信息，并举例说明DC中实体是如何转化为PROV形式，然后根据不同用户对DC与PROV之间映射复杂程度的兴趣不同，分为直接映射、复杂映射，并说明有些元素排除在映射外的原理。

3.1 DC术语中的溯源信息

根据数据溯源模型对于元数据的要求，许多DC术语可用于描述资源所需的溯源信息，包括过去何时被影响（When），谁影响它（Who）以及它如何被影响（How）。其余的DC术语则告诉我们什么被影响（What）。表2对DC术语进行分类。

每个类别对应回答DC是否包含数据溯源所需元数据项，What类别包含DC中描述性元数据而没有涉及溯源信息的术语，如dct:title、dct:abstract等。Who类别包含溯源中Agent（代理）术语。Agent指通过活动来控制实体的类。dct:contributor、dct:creator、dct:publisher通过编辑、创建、发布等动作来操控实体。dct:rightsHolder虽并没有通过活动来控制实体，但所有权者在许多领域如艺术品、图书馆领域是重要的溯源信息。因此，DC四个术语对应于数据溯源中的Agent类。When类别包含与日期、时间有关的术语。当追踪资源是何时被创建（dct:created）、修改（dct:modified）、发布（dct:issued）时，日期是有效的溯源信息记录，而可用性（dct:available）和有效（dct:valid）被认为是关于溯源信息的特殊记录。How类别包含衍生（derivation）相关术语。当一个资源衍生自其他资源时，原始资源就变成衍生资源的一部分。在DC中，衍生可被进一步分为版本（dct:isVersionOf）、格式序列化（dct:isFormatOf）、代替（dct:replaces）以及信息来源（dct:source）。dct:references是弱关系，它虽然和资源有关，但并不总意味着内容是基于它的，不过可以假设一个引用资源影响着被描述的资源，因此它和溯源有关。

3.2 DC转化为PROV实例

DC关注描述资源信息，而PROV关注资源的状态和变化，在实际应用中两者之间需要转化。以下实例1用DC语句描述一个文档（ex:prov-dc-20130312），并用Turtle格式描述了一个简单的元数据记录。

在实例1中dct:title、dct:subject用于描述资源ex:prov-dc-20130312，但它们并不提供任何关于该资源

表2 DC术语分类（属性）[17]

类别	子类别	术语（Term）
无对应溯源信息	What	abstract, accrualMethod, accrualPeriodicity, accrualPolicy, alternative, audience, bibliographicCitation, conformsTo, coverage, description, educationLevel, extent, format, hasPart, isPartOf, identifier, instructionalMethod, isRequiredBy, language, mediator, medium, relation, requires, spatial, subject, tableOfContents, temporal, title, type
溯源信息	Who	contributor, creator, publisher, rightsHolder
溯源信息	When	available, created, date, dateAccepted, dateCopyrighted, dateSubmitted, issued, modified, valid
溯源信息	How	accessRights, hasFormat, hasVersion, isFormatOf, isVersionOf, license, isReferencedBy, isReplacedBy, references, replaces, rights, source

实例1：

```
ex:prov-dc-20130312
    dct:title "A mapping from Dublin Core...";
    dct:creator ex:kai, ex:daniel, ex:simon, ex:michael;
    dct:created "2012-02-28";
    dct:publisher ex:w3c;
    dct:issued "2012-02-29";
    dct:subject ex:dublincore;
    dct:replaces ex:prov-dc-20121211.
```

是如何被创建或修改的信息。而一些描述语句涉及溯源信息，如dct:creator意味着文档被创建并涉及作者，同样，dct:issued意味着文档被发布。dct:replaces意味着将ex:prov-dc-20121211与ex:prov-dc-20130312关联起来。但是DC描述的是静态结果文档，并不能明确描述该文档的不同状态。如一个文档可能涉及dct:created、dct:issued日期，但并不能动态描述该活动产生的关联性文档。按照PROV-O，发布活动可能涉及两个不同文档的状态：发表之前和之后。一般的处理方法是创建与原始文档相关的新的实体，并用prov:specializationOf来将二者关联。如图2，DC中的publisher意味着有publish（发布）活动，用空白节点表示资源的每个状态。该活动与Agent关联，将ex:doc1具体化为文档（:_usedEntity），并产生结果文档（:_resultingEntity）。

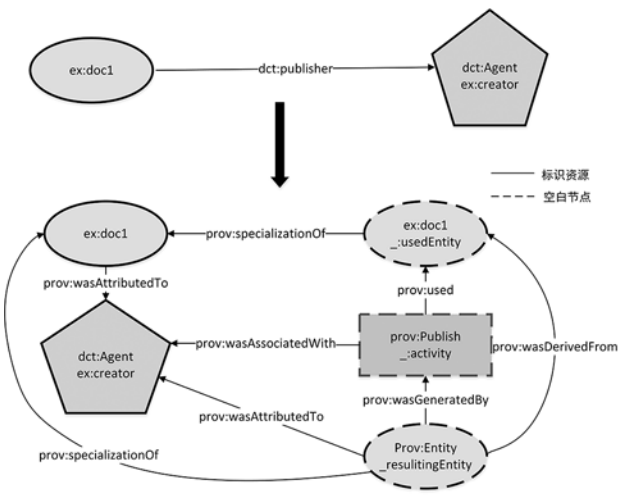

图2 创建空白节点将DC转化为PROV

3.3 直接映射

直接映射是使用RDF推理机制将DC术语直接映射到PROV二元关系中。从建设数据溯源模型角度看，DC虽不那么复杂，但它更具体的是描述关于发生活动的类型。表3、表4提供详细的DC属性和类的映射。

表3 直接映射（属性）

DC 术语	关系	PROV 术语
dct:created	rdfs:subPropertyOf	prov: generatedAtTime
dct:modified		
dct:dateAccepted		
dct:dateCopyrighted		
dct:dateSubmitted		
dct:issued		
dct:hasFormat	rdfs:subPropertyOf	prov:alternateOf
dct:isFormatOf		
dct:creator	rdfs:subPropertyOf	prov: wasAttributedTo
dct:contributor		
dct:publisher		
dct:rightsHolder		
dct:isFormatOf	rdfs:subPropertyOf	prov: wasDerivedFrom
dct:references		
dct:source		

表4 直接映射（类）

DC 术语	关系	PROV 术语
dct:Agent	owl:equivalentClass	prov:Agent
dct:BibliographicResource	rdfs:subClassOf	prov:Entity
dct:LicenseDocument		
dct:RightsStatement		
dct:PhysicalResource		
dct:LinguisticSystem	rdfs:subClassOf	prov:Plan
dct:MethodOfAccrual		
dct:MethodOfInstruction		
dct:Policy		
dct:Location	owl:equivalentClass	prov:Location
dct:ProvenanceStatement	rdfs:subClassOf	prov:Bundle

另外，dct:source和dct:isVersionOf分别是prov:hadPrimarySource和prov:wasRevisionOf的父属性，dct:LocationPeriodOrJurisdiction是prov:Location的父类。图3展示了PROV与DC之间的映射关系。

3.4 复杂映射

复杂映射包含一组定义了从DC语句到PROV语句的模式。这种模式并不总是被需要，而是用户根据实例选择是否使用它们。同时，并不是所有的直接映射都关联一个复杂的映射，而是隐含着一些特定活动，如在创建、发布时进行关联。复杂映射可通过SPARQL结构查询来提供。根据查询类别不同划分为以下三种映射：Entity-Agent映射（Who）、Entity-Date映射（When）和Entity-Entity映射（How）。在Entity-Agent映射（Who）中，一个creator（创建者）控制prov:Create活动，该活动产生了实体（?document）。Agent（代理）所扮演的角色是创建者角色（prov:Creator）。该复杂映射可通过SPARQL中的CONSTRUCT结构将DC中的术语与PROV术语进行关联，并通过WHERE语句进行查询。?document和?agent可根据具体情况来进行取值。CONSTRUCT结构可看作一个可以往里面填数值的模板。在Entity-Agent映射中，dct:contributor、dct:publisher与dct:rightsHolder的映射表达类似，只需要做小部分改变。

3.5 排除映射

因为有些属性和类不适用或并没有描述溯源信息，而被排除在映射之外。另外，有些是描述性元数据但并不描述溯源信息的类，也不把它映射为prov:Entity的子类，如dct:MediaType、dct:Standard。表5列出排除映射的相关术语。值得注意的是，available虽描述资源何时可用，但在PROV中不能与generation和invalidation概念直接映射，故将其排除在映射外。

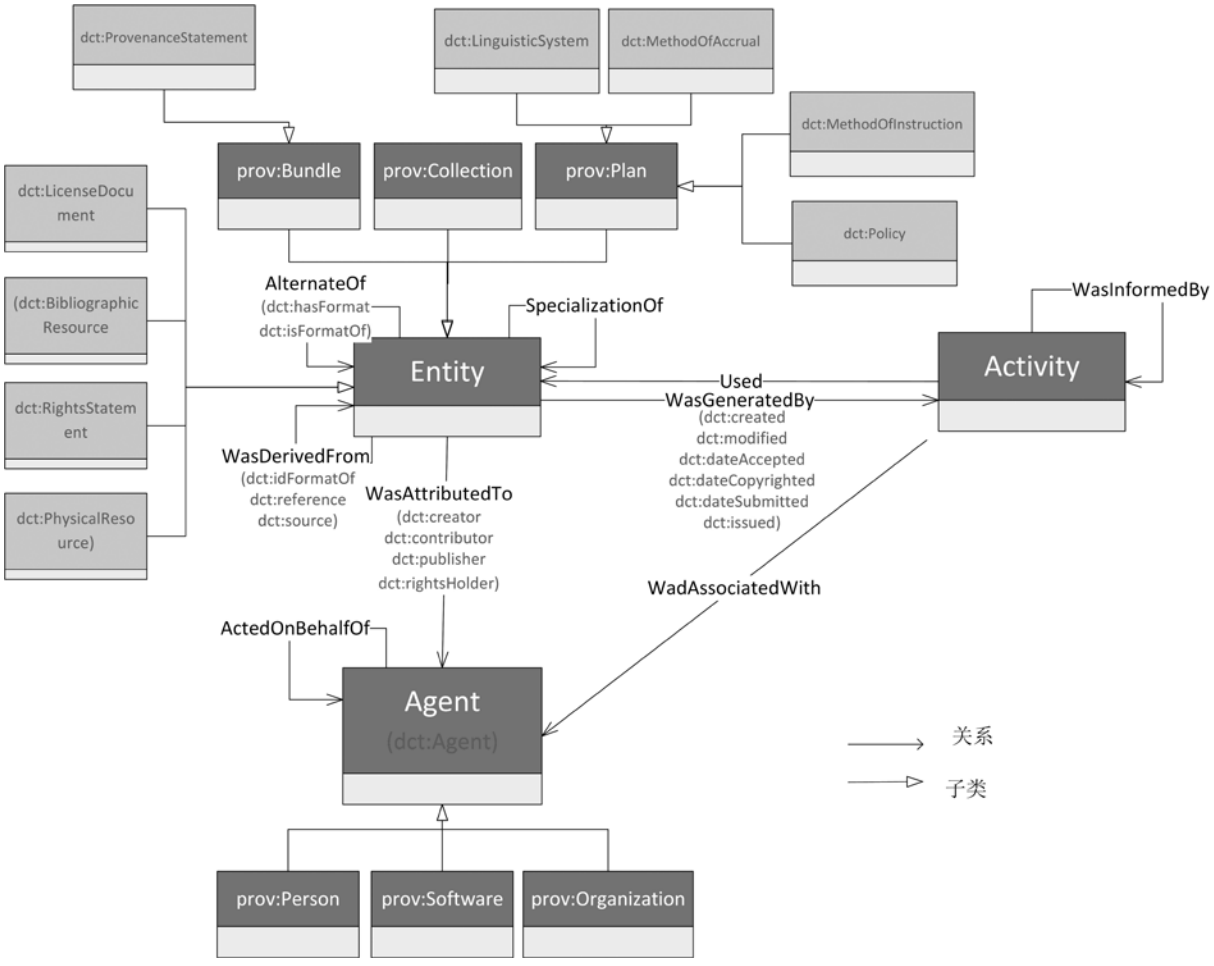

图3 PROV与DC映射

表5 排除映射的类和属性

类别	术语
排除映射的类	AgentClass、FileFormat、Frequency、Jurisdiction、MediaTypeOrExtent、PeriodOfTime、PhysicalMedium、SizeOrDuration、Standard
排除映射的属性	abstract、accessRights、accrualMethod、accrualPeriodicity、accrualPolicy、alternative、audience、available、bibliographicCitation、conformsTo、coverage、description、educationLevel、extent、format、hasPart、identifier、instructionalMethod、isPartOf、isRequiredBy、language、license、mediator、medium、rights、relation、requires、spatial、subject、tableOfContents、temporal、title、type、valid

4 基于DC元数据实现数据溯源的建议

Ram[18]数据溯源信息应包含Who、When、Where、How、Which、What、Why七部分信息。DC元数据虽涵盖了What、Who、When、How、Where、Why六方面的描述，若要满足数据溯源要求，仍需补充描述活动、关系、代理等三方面的字段。具体建议如下：

（1）补充描述活动字段

实体如何（How）演变的描述是数据溯源在数字图书馆的核心，包括描述数字对象是如何经由产生、转换、修改等系列活动而呈现最终的状态。结合表2与图3所示，虽DC元数据部分包含实体如何（How）演变字段，但描述的是实体间衍生关系和格式转换关系，如版本更新（dct:isVersionOf）、参考引用（dct:reference）、主要来源（dct:source），需要更多描述动态过程字段的补充。如DC元数据中存在描述创建者字段dct:creator以及创建时间字段dct:created，则必然存在创建活动，增加活动字段使演变过程变得清晰。建议补充描述活动字段，如dct:create、dct:modify、dct:publish等。

（2）补充描述关系字段

数据溯源实现原理在于展示实体（Entity）、活动（Activity）、代理（Agent）三个节点之间的因果关系，谁（Agent）对实体（Entity）做了些什么（Activity）。从数据溯源角度，DC元数据需要补充描述关系字段，从而将三者进行关联。建议补充描述关系字段，如dct:used、dct:wasGeneratedBy、dct:wasAssociztedWith等。

（3）补充描述代理（Agent）字段

DC元数据存在相关人员（Who）的描述字段，包括dct:creator、dct:contributor、dct:publisher、dct:rightsHolder。但管理者或用户对于数据质量的判断，除了对相关人员操作过程进行判断外，还包括对其操作人员所代表的企业、组织机构及操作人员所使用工具（Which）的认可度。建议补充描述代理字段，如dct:organization、dct:software。

5 结语

数据溯源能加强数字图书馆对数字资源的质量管理，有效促进数字资源的开发和利用。PROV与DC间的映射关系，明确了数字图书馆中包含的数据溯源信息，使研究人员能够利用DC元数据建立该领域的数据溯源技术体系，是数字图书馆领域实现数据溯源的基础和关键。本文创新点在于对PROV和DC之间的映射做了系统研究，画出PROV与DC的映射图，并建议现有DC元数据补充相关字段以实现数据溯源。今后拟在本文研究基础上将PROV的数据溯源模型在图情领域进行实证。

参考文献

[1] Chiticariu L, Tan W C, Vijayvargiya G. DBNotes: a post-it system for relational databases based on provenance [C]// Proc of the ACM SIGMOD International Conference on Management of Data, New York: ACM Press, 2005: 942-944.

[2] Fan Hao. Tracing data lineage using automed schema transformation pathways [C]// Proc of the 19th British National Conference on Databases, Berlin: Springer, 2002: 44-55.

[3] Specication C, Cliord B, Freire J, et al. The Open Provenance Model [J]. Future Generation Computer Systems, 2008, 27(6): 743-756.

[4] Provenir Ontology [EB/OL]. [2015-11-27]. http://wiki.knoesis.org/index.php/Provenir_Ontology.

[5] PROV-DM: The PROV Data Model [EB/OL]. [2015-11-27]. http://www.w3.org/TR/2013/REC-prov-dm-20130430/.

[6] 李秀美,王凤英.数据起源安全模型研究[J].山东理工大学学报(自然科

学版),2010,24(4):56-60,64.

[7] 陈颖.一种基于DNA双螺旋结构的数据起源模型[J].现代图书情报技术,2008(10):11-15.

[8] Music Ontology to Media Value Chain Ontology and PROV-O Ontology Mapping [EB/OL]. [2015-11-27]. http://oeg-dev.dia.fi.upm.es/mvco-prov/.

[9] 邓仲华,魏银珍.面向数据发布的科学工作流数据溯源方法研究[J].图书与情报,2014(3):61-66.

[10] 李文燕,吴振新.起源信息模型及标准PROV的研究分析[J].情报理论与实践,2015,38(4):23-29.

[11] 倪静,孟宪学.PROV数据溯源模型及Web应用[J].图书情报工作,2014,58(3):13-19.

[12] 倪静,孟宪学.Web应用中起源信息的定位和查询机制研究[J].图书情报工作,2014,58(11):97-103.

[13] 吴振新,李文燕.起源技术在长期保存中的应用与研究[J].图书情报工作,2015,59(8):118-125.

[14] 李秀丽,徐越权.浅析DC在高校图书馆网络信息资源组织中的应用[J].农业图书情报学刊,2010,22(5):111-114.

[15] Provenance-Overview [EB/OL]. [2015-12-09]. http://www.w3.org/TR/2013/NOTE-prov-overview-20130430/.

[16] DCMI Metadata Provenance Task Group [EB/OL]. [2015-12-09]. http://dublincore.org/groups/provenance/.

[17] Dublin Core to PROV Mapping [EB/OL]. [2016-03-17]. http://www.w3.org/TR/2013/NOTE-prov-dc-20130430/.

[18] Ram S, Liu J. A new perspective on semantics of data provenance [EB/OL]. [2016-03-18]. http://ceur-ws.org/Vol-526/InvitedPaper_1.pdf.

作者简介

林芳芳，女，1992年生，中国科学技术信息研究所硕士研究生，研究方向：信息资源管理，E-mail：linff2014@istic.ac.cn。
赵辉，女，1971年生，中国科学技术信息研究所副研究馆员，研究方向：信息资源管理、科技资源管理。

Study on Mapping between Data Provenance Model and DC Metadata

LIN FangFang, ZHAO Hui
(Institute of Scientific and Technical Information of China, Beijing 100038, China)

Abstract: PROV is a series of standard about data provenance, which is closely related with the DC metadata. In the field of digital library, the traceability data become a kind of trend, and how to use DC metadata to establish data provenance system has become one of the key research. This paper starts with the concept of DC metadata and PROV, compares and analyses the relationship between them, and discusses the mapping between them. It comes to a conclusion that if we want to meet the data provenance requirement, DC needs to add the description about activity, relationship and agent.

Keywords: Data Provenance; PROV; DC Metadata; Mapping; Digital Library

（收稿日期：2016-01-18）

书讯

《网络环境下叙词表编制与发展》

叙词表作为一种有效的知识组织工具，在网络环境下继续发挥着重要作用。中国科学技术信息研究所常春研究馆员及其项目研究团队，依托国家社科基金项目“网络环境下叙词表的编制模式与应用方式研究”（10BTQ048），对这一课题开展了研究，并于近期完成了《网络环境下叙词表编制与发展》一书。

该著作主要论述了网络环境下叙词表的编制、维护与应用的理论和方法。编制方法包括网络环境下总体策略、总体形态、选词方法、词间关系建立方法、编制管理机制、维护方式方法等；应用研究包括网络环境下相关技术的突破给叙词表带来的各类新的应用方式，从术语服务、多语种翻译、概念组配、知识单元、概念映射、国外应用等多个方面，阐述了网络环境下叙词表的发展方向。最后按年代顺序介绍了国内历年编制的、可查阅的重要中文叙词表，理、工、农、医四大领域20多个可从网络上在线获取的英文叙词表。可供图书馆学、情报学专业相关专业人员参考使用。

《网络环境下叙词表编制与发展》于2015年4月由科学技术文献出版社出版，定价38.00元。