数字图书馆

语义网环境下数据溯源表达模型研究综述*

沈志宏[1,2,3] 张晓林[1]

[1](中国科学院国家科学图书馆 北京 100190)

[2](中国科学院计算机网络信息中心 北京 100190)

[3](中国科学院研究生院 北京 100049)

【摘要】综述语义网环境下数据溯源在表达模型与技术上的研究进展,重点研究 Open Provenance Model、Provenir Ontology 与 Provenance Vocabulary 的描述方法和能力,结合科研环境,讨论这些溯源模型在使用和推广上所面临的挑战。

【关键词】数据溯源 工作流溯源 溯源模型 开放溯源模型 语义网 关联数据

【分类号】TP393

Data Provenance Model in Semantic Web Environment: An Overview

Shen Zhihong[1,2,3] Zhang Xiaolin[1]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](Computer Network Information Center, Chinese Academy of Sciences, Beijing 100190, China)

[3](Graduate University of Chinese Academy of Sciences, Beijing 100049, China)

【Abstract】This paper reviews the progress of research on provenance in the Semantic Web environment, and introduces provenance models including the Open Provenance Model, Provenir Ontology and Provenance Vocabulary, focusing on the description methods and description capabilities of them. Finally, it discusses the difficulties and new challenges when applying these provenance models in the scientific research environment.

【Keywords】Data provenance Workflow provenance Provenance model Open provenance model Semantic Web Linked data

1 概 述

在当今开放的网络环境中,人们常会在 Web 上发现一些有问题的、甚至自相矛盾的信息。判定这些信息的真实性往往需要借助于 Data Provenance。Provenance 翻译成“溯源”、“起源”,数据溯源也称为“数据族系(Data Lineage)”、“数据系谱(Data Pedigree)”、“数据来源(Data Derivation)”等。通过溯源,人们可以根据艺术品的出处和所有者来鉴别艺术品的真实性,了解食品工业、汽车行业中产品的生产流程,明白科研活动中的数据是如何基于科学工作流计算出来的还是基于传感器、仪器采集而来的。在分布式网络环境(如:e-Science,Web of Linked Data)中,很多数据驱动的应用都会集成和融合一些数据,如果不记录这些原始数据的溯源信息,融合后的数据其真实性和有效

收稿日期:2011-03-23

收修改稿日期:2011-04-06

*本文系中国科学院信息化专项基金“数据应用环境建设与服务”(项目编号:INFO-115-C01)和国家科技基础条件平台建设项目“基础科学数据共享网—理化天文空间生物”课题基金“标准规范及共享服务平台建设”(项目编号:BSDN2009-17)的研究成果之一。

性将会有所降低。

国际上讨论溯源技术的组织和会议很多,比较早的包括 Data Provenance/Derivation Workshop[1]、Data Provenance and Annotation[2]、Provenance Aware Storage Systems(PASS)[3]等,较近的包括 International Provenance and Annotation Workshop Series(IPAW)[4]、Workshop on the Theory and Practice of Provenance(TaPP)[5]、Principles of Provenance (PrOPr)[6]、Provenance in Practice Workshop(PPW)[7]、International Workshop on the Role of Semantic Web in Provenance Management(SWPM)[8]、International Workshop on Data and Process Provenance(WDPP)[9]等。另外,W3C 也于 2009 年 9 月专门开设了 W3C Provenance Incubator Group[10](以下简称 PROV - XG)来研究语义网环境下的溯源。

文献[11]对溯源相关的文献做过统计,结果如图 1 所示:

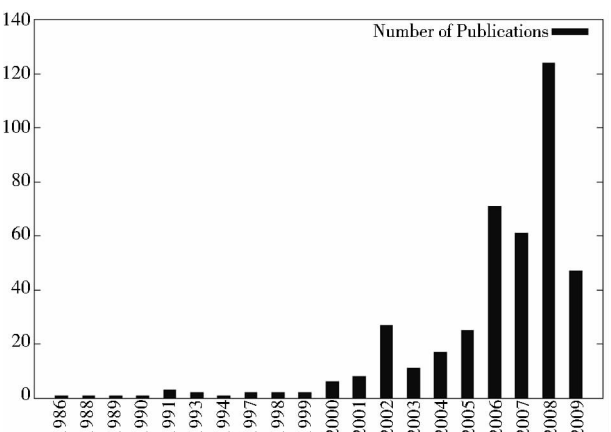

图 1 溯源技术研究文献统计图[11]

从 1986 年到 2009 年,关于溯源的发表物为 425 篇,最早的文献[12]可以追溯到 1986 年,它介绍了一项用以帮助分析人员理解和校验数据结果的审计技术。通过图 1 可以发现,基本上有一半的文献都是 2006 年以后发表的。另外,不难看出文献发表的几次高峰,这与溯源研究社区的活动也正好一致:2002 年,由 Foster 与 Buneman 组织了第一次溯源工作组会议;2006 年,由 Foster 和 Moreau 组织了第一届 IPAW 会议[13];2008 年,由 Freire 和 Moreau 组织了第二届 IPAW 会议[14]。另外,2010 年 6 月,IPAW2010[15]在纽约召开,该次会议发表了各种文章和报告共 41 篇,关注的主题包括溯源模型、溯源架构与工具、溯源数据的安全保障、Linked Data 下溯源信息的发布和消费等。

国内针对溯源的研究不多,戴超凡等[16]对数据溯源的发展现状做了综述,并对数据仓库中数据日志跟踪的理论与方法进行研究。李亚子[17]对数据溯源标注模式与描述模型做了综述,陈颖[18]结合目前的描述模型提出了一种基于 DNA 双螺旋结构的数据溯源模型。此外,乐鹏等[19]结合空间数据研究了基于 SOA 的空间数据溯源,李秀美等[20]则从安全的角度研究了数据溯源的安全模型。

随着语义网的发展和溯源研究的不断深入,人们越来越迫切地需要实现不同溯源系统的信息交换与互操作,这催生了各种形式化、语义化的溯源表达模型的产生,本文就语义网环境下数据溯源在表达模型与技术上的研究进展进行综述。

2 数据溯源的概念与定义

溯源的概念由来已久,但直到 2007 年 Buneman 等[21]才明确地提出将溯源划分成两类,即粗粒度的工作流溯源(Workflow Provenance)和细粒度的数据溯源(Data Provenance)。

数据溯源指某个转换步骤结果中的片段数据(Single Pieces of Data)是如何衍生的。最传统的例子是针对数据库 D 通过一个 SQL 查询得到结果Q(D)。2001 年 Buneman 等[22]将数据库领域中的溯源(数据溯源)区分为 Why - provenance 和 Where - provenance,2007 年 Green 等[23]在此基础上引入了 How - provenance。2010 年 Ram 等[24]提出 W7 模型,指出数据溯源信息应该包括 Who、When、Where、How、Which、What、Why 7 个部分。工作流溯源用来记录工作流中产生最终输出的完整过程。文献[25]用工作流模型(时序图)和构建图对溯源进行描述,该模型由 4 个基本的组件组成:工作流引擎、服务、为服务提供功能的应用和在工作流里生产与消费的数据。

可以看出,人们尝试从不同的视角给溯源以不同的定义和理论模型。文献[11]对这些不同的定义给出了全面的综述。PROV - XG 给出了数据溯源的工作定义(Working Definition)[26],它认为一个资源的溯源指的是对资源的生产、传递等影响中的实体和流程的描述记录,溯源为评估真实性、增加信任、再现过程提供了必要的基础。

数据溯源根据应用领域的不同,其包含的内容也不同。本文结合 PROV - XG 给出的一些示例与其他文献[27,28],整理如表 1 所示:

表 1　数据溯源包含的内容

应用领域	数据溯源包含的内容
Web 领域	Web 资源的创建、发布、访问,以及讨论、链接和复用等行为
艺术领域	艺术品的创建(谁创建? 何时何地? 为何创建? 如何创建?),以及与时间相关的描述型元数据(如:与作品相关的化学成分,对画笔风格的分析等)
科学研究领域	对样本所做的物理的、计算的处理流程,样本的描述信息,实验协议等
商务领域	财务法律流程,电子(如:在线订单)或物理(如:运输)流程等
制造行业	产品的设计、制造流程等
图书馆领域	数字对象生产、转换、修改等过程的信息等

溯源与元数据之间存在着某种重合的关系,溯源信息实际上是一种上下文的元数据。PROV - XG 认为,元数据用以描述对象的属性,而对象属性往往包含着数据溯源信息,因此两者在有些情况下是一致的。图书馆界在考虑数字资源长期保存时,在长期保存元数据中也考虑了溯源。例如,CEDARS(CURL Exemplars in Digital ARchiveS)[29]将保护描述信息细分为确认信息、环境信息、固化信息以及溯源信息。另外,根据 NSF[30]在 2008 年对元数据的定义"元数据为数据的子集,用以概括数据的内容、环境、结构、相互关系和溯源"也可以看出,数据溯源信息可以认为是元数据的一部分。以一条科学数据为例,它往往会包含创建者和创建时间等元数据项,当从溯源的角度来分析时,每次所做的创建、修改的过程以及涉及到的一些主体(人员或者程序)、时间、空间、生成方法(如:根据影像修复算法),这些元数据项就属于溯源信息。

溯源与信任(Trust)也存在着一定的关系,通常认为,可以从完整的溯源信息中推导出信任。信任是一个基于上下文的主观性判断,而溯源为信任的推导提供客观的记录。当然,在推导的过程中,对主体的身份认证尤显重要。但是身份认证在分布式、异域环境下会显得格外复杂,如:用户的唯一标识、数字签名、访问控制策略等。

3　数据溯源表达模型

得益于语义网的发展,在 2006 年芝加哥召开的第一届 IPAW 会议[13]中的一个讨论溯源标准化的分会上,溯源研究社区提出了需要更好地理解不同溯源系统的性能和表达,并研究它们的异同与设计动机;同时期望通过溯源信息的交换,建立起系统之间的互操作。2007 年 8 月,继两届溯源竞坛(Provenance Challenge)之后,在盐湖城的一个工作组会议之后,Moreau 等[31]发布了开放溯源模型(Open Provenance Model,OPM)。此后,关于数据溯源的表达模型竞相开发出来。除了 OPM 之外,还有 Provenir Ontology、Provenance Vocabulary、SWAN Provenance Ontology 等。此外还有已经广为接受的词表,如:Proof Markup Language、Dublin Core、PREMIS、WOT Schema、Semantic Web Publishing Vocabulary、Changeset Vocabulary,尽管它们不是专门针对数据溯源而提出的,但是与溯源模型存在着一些交叉和重合。因此,人们常常研究这些模型之间的异同与映射技术。

3.1　开放数据溯源模型 OPM

OPM 规范 1.0 版本于 2007 年 8 月发布。随着几届溯源竞坛的相继展开,OPM 规范推出 1.01 版本、1.1 版本。OPM 设计的目标是为不同的系统提供可交换的溯源信息,并允许开发人员创建并共享操作该模型的工具。OPM 同时从技术角度定义了溯源,支持对任何事物(不仅仅是针对计算机系统)的溯源,并允许多级描述同时共存。

OPM 首先定义了三个核心概念,即 Artifact、Process 和 Agent。Artifact 用以指代一个状态,它可以是物理的一个对象,也可以是计算机系统中的一个数字化表达。Process 指代由 Artifact 引起的一个或者一系列的动作。Agent 指代 Process 的催化剂,它用以促进、控制和影响 Process 的执行。此外,OPM 还引入了 Role 的概念,一个 Process 可能会产生多个 Artifact,这些 Artifact 就会拥有不同的 Role。以某次除法运算为例,Agent 为计算器(或者运算程序),Process 为除法运算,参与运算的有两个 Artifact,它们分别属于除数和被除数这两个角色,运算的结果也包含两个 Artifact,它们则分别属于商和余数这两个角色。

OPM 常常采用有向无环图来表示溯源图,Artifact、Process 和 Agent 表示为节点,它们之间的关联表示为边。图 2 列出了 OPM 各节点之间所有可能的关联[32]。可以看出,同一类节点之间也会存在关联,如:某个 Artifact 可能会 wasDerivedFrom 另一个 Artifact,某个 Process 会 wasTriggeredBy 另外一个 Process。文献[32]为这些关系分别给出了严格的定义,如:P2 wasTriggeredBy P1 被定义成"启动 P1 为完成 P2 的必要条件"。

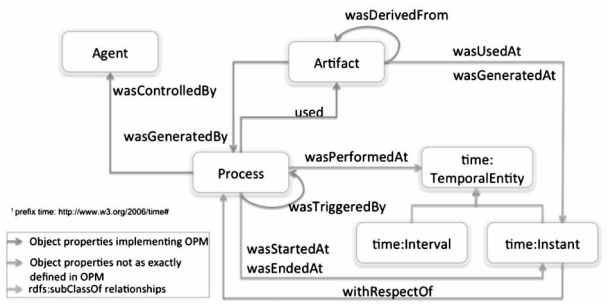

图 2　OPM 溯源模型结构图[32]

绘制 OPM 溯源图时，一般使用椭圆表示 Artifact，矩形表示 Process，八角菱形表示 Agent。如：John 烤蛋糕的过程溯源可以描述成如图 3 所示：

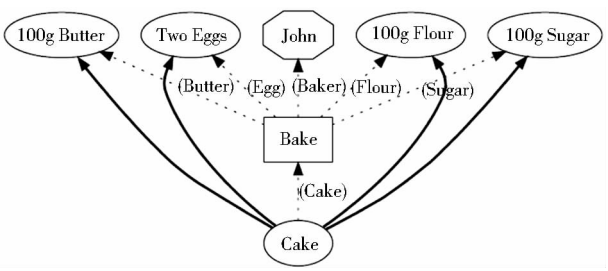

图 3　John 烤蛋糕的过程溯源[32]

3.2　Provenir 数据溯源模型

2008 年，美国赖特大学 Sahoo 等[33]在第二届 IPAW 会议上提出了 Provenir 模型，Provenir 来自于法语，意思为“to come from”。

Provenir 数据溯源模型给出了 Provenir Ontology[34]。Provenir Ontology 定义了三个主要的类作为模型的基本组件，它们是 Data、Process、Agent。Data 类代表科学实验中的原始材料、中间材料、最终产品以及影响科学流程执行的一些参数。Process 和 Agent 的含义与 OPM 中的相似。不过 Provenir Ontology 强调了两个概念，即 Occurrent 和 Continuant，Occurrent 指那些随着时间变化而变化的偶然性的特性，Continuant 相反，指那些不随时间变化而改变的持续性的特性。Provenir Ontology 认为，Process 是 Occurrent 的，Data 和 Agent 是 Continuant 的。

Data 类具有两个子类 Data_collection 和 Parameter，Data_collection 代表科学过程中参与的数据实体，Parameter为影响科学过程的一些个体，它又可细分为三个子类 Temporal_parameter、Spatial_parameter、Domain_parameter，分别代表时间的、空间的、领域相关的参数。Provenir Ontology 类的关系如图 4 所示[33]：

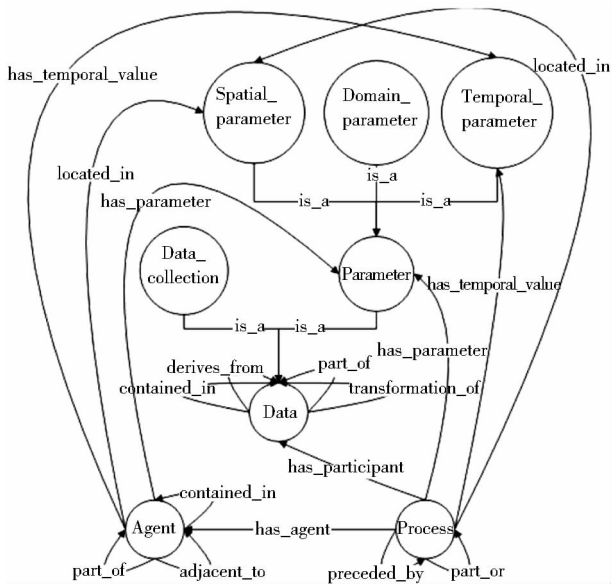

图 4　Provenir Ontology 模型结构图[33]

文献[33]还对常见的溯源信息查询模式进行了总结，并针对 Provenir Ontology 给出了查询函数，同时一一给出了形式化定义。笔者整理如表 2 所示：

表 2　溯源信息查询函数定义

查询类别	查询函数名称	含义	示例
查询溯源元数据	Provenance( )	返回给定数据实体的所有溯源元数据	返回 HyperCube85357162234026 所有的溯源元数据
	Provenance _ pathway( )	仅返回工作流相关的(实例为 Data _ collection 和 Process 类型)溯源元数据	返回 2003－4－1 到 2003－5－2 之间的洋流流程相关的溯源元数据
查询数据值	Provenance _ context( )	返回满足给定溯源元数据约束的所有 Data_collection	返回标识为 oceanBuoy7044 的海洋浮标 2003－4－1 到 2003－5－2 之间处于损坏状态的所有数据集
	Pc_process( )	返回满足给定溯源元数据约束的所有 Process	返回以 InverseData 为 False 的 HyperCubetoDataTable 计算过程的所有调用
	Pc_agent ( )	返回满足给定溯源元数据约束的所有 Agent	
修改溯源元数据	Provenance _ compare( )	比较两部分溯源信息	比较两幅海洋可视化图形的溯源信息(传感器的类型、科学工作流的输入参数等)
	Provenance_merge ( )	合并两部分溯源元数据	

Provenir Ontology 在不同领域中都得到了应用，包括生物医学科学、海洋学、传感器、卫生保健等领域。

3.3　Provenance Vocabulary

在 LDOW2009(Linked Data On the Web)工作组会议[35]上，Hartig 等[36]提出了一个面向 Web 上 Linked Data 的溯源模型，并给出其 OWL 本体模型 Provenance Vocabulary。Provenance Vocabulary 从两个维度描述数据溯源，即数据创建(Data Creation)与数据访问(Data

Access)。数据访问在 Web 上很常见,但是在此之前,人们很少关注描述数据访问过程中的溯源。

Provenance Vocabulary 定义了一个核心的词表(http://purl.org/net/provenance/ns#),并提供了三个扩展模块:Types(http://purl.org/net/provenance/types#)、Files(http://purl.org/net/provenance/files#)和 Integrity Verification(http://purl.org/net/provenance/integrity#)。Provenance Vocabulary 包括通用词汇、用于数据创建的词汇以及用于数据访问的词汇。与前两个模型类似,通用词汇定义了三个核心概念:Actor、Execution 和 Artifact。Actor 具有两个子类 HumanActors 和 NonHumanActors,HumanActors 用以指代人、组织或者公司,NonHumanActors 则指代软件、工具、算法等。Artifact 又细分为 DataItems 和 Files 两个子类。Provenance Vocabulary 的结构如图 5 所示[37]:

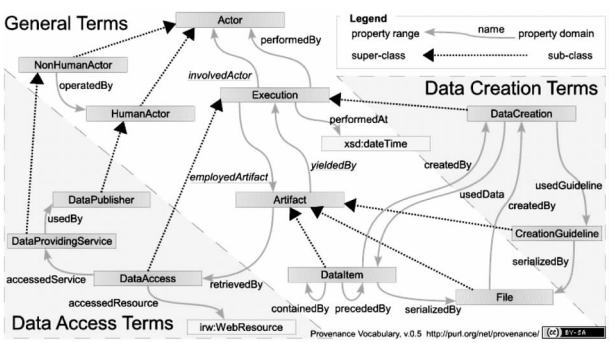

图 5　Provenance Vocabulary 模型结构图[37]

Provenance Vocabulary 可以描述数据创建的不同方式,文献[37]中给出两个例子,如:Alice 在 2009 年 7 月 10 日手动创建了一个 RDF 文件,可以描述如下:

```
< > rdf:type prv:DataItem;
  rdf:type <http://www.w3.org/2004/03/trix/rdfg-1/Graph>;
  prv:createdBy [ rdf:type prv:DataCreation;
        prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime;
        prv:performedBy <http://example.org/Alice> ].
```

另外一个例子是描述 Bob 通过传感器 Sensor1 采集到一条数据,描述如下:

```
_:a rdf:type prv:DataItem;
  prv:createdBy [ rdf:type prv:DataCreation, prvTypes:Measurement;
      prv:performedAt "2009-07-10T12:00:00Z"^^xsd:dateTime;
      prv:performedBy <http://example.org/Sensor1> ].
<http://example.org/Sensor1> rdf:type prv:Actor, prvTypes:Sensor;
      prv:operatedBy <http://example.org/Bob>.
```

除了以上介绍的三种溯源模型之外,在其他一些广为接纳的模型中其实也包含溯源描述的能力。如:与 OPM 相比,Dublin Core 元数据词汇[38]所包含的 dct:Agent、dct:Source、dct:Contributor、dcmitype:Event、dct:ProvenanceStatement 等都具有可映射的溯源信息。另外还有一些词汇,如 dct:Contributor、dct:Creator、dct:Publisher,以及 dct:Created、dct:dateAccepted、dct:dateCopyrighted、dct:dateSubmitted、dct:Modified 等属性名,它们与 OPM 词汇在含义上也存在着某种重合,但不存在直接的映射方法。

PROV-XG 对不同的词表做了较为细致的研究,并针对包括 Provenir Ontology、Provenance Vocabulary 等在内的 9 个词表,研究了它们与 OPM 之间的映射方法[39]。通过采用 SKOS[40]的 skos:broader、skos:narrower 和 skos:related,以及 OWL2[41]的 owl:equivalentClass、owl:equivalentProperty,给出了不同的类与属性之间的映射关系。

4　结　语

近年来随着语义网的深入,人们已经不再局限于溯源的理论研究和在单个领域中的应用,而是开始考虑如何将溯源以形式化的方式表达出来,以及如何实现不同溯源系统之间的互操作。通过对比可以看出,这些表达模型基本上都存在着一些值得借鉴的共性。

(1)这些溯源模型都没有给出一套统一的表达,而是鼓励人们针对不同应用领域对其进行继承和扩展。可以看出,无论 OPM、Provenir Ontology 还是 Provenance Vocabulary,它们首先给出的还是三要素(Artifact、Process、Agent),同时允许基于此参考模型进行扩展。以 Provenir Ontology 为例,很显然它无法给出一个包罗万象的溯源本体,因此它推荐以其为通用参考模型,结合特定领域(如生物学、海洋科学、天文学等)来构造各自的溯源本体模型,如:ProPreO 代表蛋白质组学领域特定的本体模型。

(2)尽管各种模型没有对具体的描述粒度做出严格的限制,但在表达能力上都支持不同描述粒度的细化和合并。图 6 是 OPM 给出的一个例子,左侧展示的列表(3,7)由列表(2,6)执行加 1 的操作生成,右侧展

示的则是更为细节的操作,列表(2,6)首先被分解成元素2和6,它们各自被加1,生成3和7,最后再被组装成列表(3,7)[37]。至于应该采取哪一种表达方式,取决于应用的需求。OPM甚至还提供了将左右这两种描述合并在一起的表达能力。

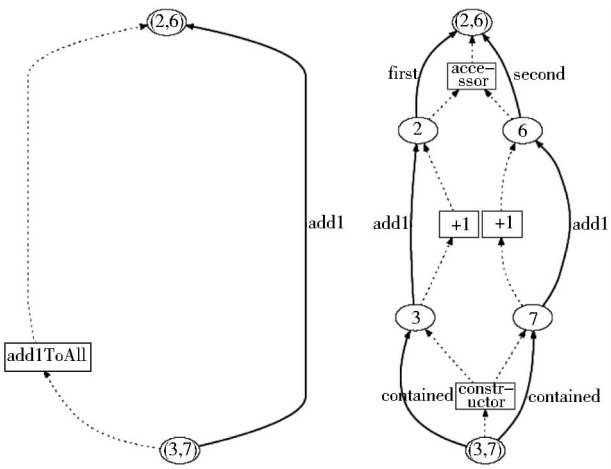

图6 溯源图的不同粒度[37]

但是目前,无论是如上这些溯源模型自身的表达能力上,还是针对它们而提出的技术实施框架上,都存在着一些急需改进和深入的地方。在真实的科研环境中,这些溯源模型的使用和推广还面临着较大的挑战。

(1)如何保证科学流程的溯源信息的完全可重算能力。

重算(再现)数据的产生过程是数据溯源的重要目标,但是在实际的科学计算中,往往会包含很大的数据量(大矩阵或者上百兆(M)、上吉(G)的数据文件),参与计算的环境极其复杂,不仅会依赖于特定的自然条件,还会依赖于特定的软件平台和环境,计算的过程也可能会涉及到多个系统之间的切换和配合,在这种情况下,如何来记录足够详尽的溯源信息,仍是一个很难解决的问题。即便有一个足够强大的溯源模型将这些上下文信息描述出来,如何构造一个重算框架,还原这些上下文,也是一个比较重要的问题。

(2)目前在Web上还缺乏一套统一的溯源信息发布与访问机制。

通过Linked Data等技术,人们可以将数据以统一的方式发布出去,但是如何再提供该数据的溯源信息,这些溯源和数据本身又该如何关联,目前还没有成熟的办法。文献[42]就如何在Web上集成数据的溯源信息作了简略的讨论,包括如何在HTTP响应头或者描述内容本身来包含溯源信息,如何以引用的方式(by Reference)或者以值的方式(by Value)来包含这些溯源信息等。溯源信息可以针对一个Web资源、一个Web资源的描述,以及一个Web资源的某个状态。那么如何来发布这些信息,又如何采用必要的策略来保障这些溯源信息自身的安全,目前还缺乏成熟的技术框架。

(3)如何创建这些溯源信息,这也是溯源研究一直关注的一个问题。

数据溯源的创建基本上可以通过两种方式:

①采用查询求逆(或者构造一个逆查询)推算出溯源信息,这种方式称为"查询反演(Query Inversion)"[43],又由于它是在需要用到数据溯源时才进行计算分析,因而又称为"Lazy"方式。

②采用标注的方法,即直接在数据上注明其来源。由于这种方式是在一开始就让数据通过标注[44]携带一些数据溯源信息,因而又称为"Eager"方式。

面向复杂科学过程的数据溯源,需要提供一种自动化的标注方法。本文认为,溯源模型的发展和标准化会带来另外一个后果,即使得溯源信息标注插件的标准化成为可能。一方面,硬件(如摄像头、传感器)、软件(如文档处理系统、数据加工软件、科学工作流系统)厂商可以结合溯源模型,提供一些可插拔的、可扩展的溯源信息标注插件接口(插槽),并定义插槽与插件之间交换的溯源信息的格式。另一方面,开发人员遵循这些协议和格式,开发出一些通用的、专用的标注插件。最终用户一旦安装了这些插件,即可完成标准化的溯源信息的自动创建。

以上这些困难会在相当长的时间内继续存在,但是不可否认,借助于数据溯源,有利于评估数据质量和可靠性,查询数据来源,再现数据的产生过程,发生错误时能够快速定位产生错误的位置从而分析出错误原因等。数据溯源无论对科研领域还是商业领域,都具有重要的意义。因此,如何提供更完善的溯源表达模型,以及基于这些模型提供更为便捷强大的解决方案,是溯源研究的下一步方向。

参考文献:

[1] Data Provenance/Derivation Workshop Position Papers and Talks

[EB/OL]. [2010 - 11 - 01]. http://people. cs. uchicago. edu/~yongzh/position_papers. html.

[2] Data Provenance and Annotation [EB/OL]. [2010 - 11 - 01]. http://www. nesc. ac. uk/esi/events/304/.

[3] Provenance Aware Storage Systems [EB/OL]. [2010 - 11 - 01]. http://www. eecs. harvard. edu/~margo/pass/fall - 2005. html.

[4] International Provenance and Annotation Workshop Series [EB/OL]. [2010 - 11 - 01]. http://www. ipaw. info/.

[5] 1st Workshop on the Theory and Practice of Provenance (TaPP'09) [EB/OL]. [2010 - 11 - 01]. http://www. usenix. org/events/tapp09/.

[6] Workshop on Principles of Provenance (PrOPr) [EB/OL]. [2010 - 11 - 01]. http://wiki. esi. ac. uk/PrinciplesOfProvenanceWorkshop.

[7] Provenance in Practice Workshop 2009 (PPW'09) [EB/OL]. [2010 - 11 - 01]. http://stan. cc. swin. edu. au/~samar/ppw09/.

[8] SWPM - 2010 - Knoesis Wiki [EB/OL]. [2010 - 11 - 01]. http://wiki. knoesis. org/index. php/SWPM - 2010.

[9] International Workshop on Data and Process Provenance [EB/OL]. [2010 - 11 - 01]. http://itee. uq. edu. au/~dasfaa/workshop/wdpp/WDPP09. htm.

[10] W3C Provenance Incubator Group Wiki [EB/OL]. [2010 - 11 - 01]. http://www. w3. org/2005/Incubator/prov/wiki/Main_Page.

[11] Moreau L. The Foundations for Provenance on the Web[J]. *Foundations and Trends in Web Science*, 2010, 2(2 - 3):99 - 241.

[12] Becker R A, Chambers J M. Auditing of Data Analysis[J]. *SIAM Journal on Scientific and Statistical Computing*, 1988, 9(4):747 - 760.

[13] IPAW'06 - International Provenance and Annotation Workshop [EB/OL]. [2010 - 11 - 01]. http://www. ipaw. info/ipaw06/index. html.

[14] IPAW 2008 Second International Provenance and Annotation Workshop [EB/OL]. [2010 - 11 - 01]. http://www. sci. utah. edu/ipaw2008/index. html.

[15] The Third International Provenance and Annotation Workshop (IPAW 2010) [EB/OL]. [2010 - 11 - 01]. http://tw. rpi. edu/portal/IPAW2010.

[16] 戴超凡, 王涛, 张鹏程. 数据起源技术发展研究综述[J]. *计算机应用研究*, 2010,27(9):3215 - 3221.

[17] 李亚子. 数据起源标注模式与描述模型[J]. *现代图书情报技术*,2007(7): 10 - 13.

[18] 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.

[19] 乐鹏, 彭飞飞, 龚健雅. 基于 SOA 的空间数据起源研究[J]. *地理与地理信息科学*, 2010, 26(3):6 - 10.

[20] 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.

[21] Buneman P, Tan W C. Provenance in Databases[C]. In: *Proceedings of the* 2007 *ACM SIGMOD International Conference on Management of Data*, Beijing, China. New York, USA: ACM, 2007: 1171 - 1173.

[22] Buneman P, Khanna S, Wang - Chiew T. Why and Where: A Characterization of Data Provenance[C]. In: *Proceedings of ICDT* 2001. Berlin: Springer, 2001: 316 - 330.

[23] Green T J, Karvounarakis G, Tannen V. Provenance Semirings [C]. In: *Proceedings of the* 26*th ACM SIGMOD - SIGACT - SIGART Symposium on Principles of Database Systems*, Beijing, China. New York, USA:ACM, 2007: 31 - 40.

[24] Ram S, Liu J, George R T. PROMS: A System for Harvesting and Managing Data Provenance [EB/OL]. [2010 - 11 - 01]. http://kartik. eller. arizona. edu/WITS_DEMO_final. pdf.

[25] Simmhan Y L, Plale B, Gannon D. A Framework for Collecting Provenance in Data - Centric Scientific Workflows [EB/OL]. [2007 - 01 - 10]. http://www. cs. indiana. edu/dde/papers/SimmhanICWS06. pdf.

[26] What is Provenance[EB/OL]. [2010 - 11 - 01]. http://www. w3. org/2005/Incubator/prov/wiki/What_Is_Provenance.

[27] PREMIS: Preservation Metadata Maintenance Activity (Library of Congress) [EB/OL]. [2011 - 02 - 01]. http://www. loc. gov/standards/premis/index. html.

[28] 张晓林. 元数据研究与应用[M]. 北京:北京图书馆出版社, 2002:174 - 177.

[29] Day M. CEDARS: Digital Preservation and Metadata [EB/OL]. [2011 - 02 - 01]. http://www. ercim. eu/publication/ws - proceedings/DELOS6/cedars. pdf.

[30] 2007 Cyberinfrastructure Vision for 21st Century Discovery [EB/OL]. [2008 - 03 - 01]. http://www. nsf. gov/od/oci/ci_v5. pdf.

[31] The Open Provenance Model (v1. 00) [EB/OL]. [2011 - 02 - 01]. http://eprints. ecs. soton. ac. uk/14979/1/opm. pdf.

[32] Open Provenance Model Vocabulary Specification [EB/OL]. [2010 - 10 - 10]. http://open - biomed. sourceforge. net/opmv/ns. html.

[33] Sahoo S S, Barga R S, Goldstein J, et al. Provenance Algebra and Materialized View - based Provenance Management[C]. In: *Proceedings of the* 2*nd International Provenance and Annotation Workshop*. Berlin:Springer, 2008: 531 - 540.

[34] Provenir Ontology [EB/OL]. [2010 - 07 - 10]. http://wiki. knoesis. org/index. php/Provenir_Ontology.

[35] Linked Data on the Web (LDOW2009) [EB/OL]. [2010 - 11 - 01]. http://events. linkeddata. org/ldow2009/.

[36] Hartig O. Provenance Information in the Web of Data [C]. In:

Proceedings of the Linked Data on the Web (LDOW) Workshop at WWW, Madrid, Spain. 2009.

[37] Guide to the Provenance Vocabulary [EB/OL]. [2010 - 07 - 10]. http://sourceforge.net/apps/mediawiki/trdf/index.php? title = Guide_to_the_Provenance_Vocabulary.

[38] DCMI Metadata Terms [EB/OL]. [2010 - 10 - 10]. http://dublincore.org/documents/dcmi - terms/.

[39] Provenance Vocabulary Mappings [EB/OL]. [2010 - 12 - 10]. http://www.w3.org/2005/Incubator/prov/wiki/Provenance_Vocabulary_Mappings.

[40] SKOS Simple Knowledge Organization System [EB/OL]. [2010 - 05 - 10]. http://www.w3.org/2004/02/skos/.

[41] OWL 2 Web Ontology Language Document Overview [EB/OL]. [2009 - 10 - 10]. http://www.w3.org/TR/owl2 - overview/.

[42] Provenance and Web Architecture [EB/OL]. [2010 - 11 - 01]. http://www.w3.org/2005/Incubator/prov/wiki/Provenance_and_Web_Architecture.

[43] Buneman P, Khanna S, Tan W C. Data Provenance: Some Basic Issues[C]. In: *Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science*. London, UK: Springer - Verlag, 2000: 87 - 93.

[44] Buneman P, Khanna S, Tan W C. On Propagation of Deletions and Annotations Through Views[C]. In: *Proceedings of the 21st ACM SIGMOD - SIGACT - SIGART Symposium on Principles of Database Systems*, Madison, Wisconsin. New York, USA: ACM, 2002: 150 - 158.

(作者 E - mail: shenzhihong@mail.las.ac.cn)

Mellon 基金赞助 CLIR 和斯坦福大学进行关联数据的研究及研讨会

美国图书馆和信息资源委员会(CLIR)近期获得了 Andrew W. Mellon 基金会提供的 49 500 美元资助,用于对语义网、关联数据和 RDF 三元组技术有关的出版物、项目和环境进行深入调研。同时,斯坦福大学图书馆也获得了 50 000 美元的资助来组织一个邀请型研讨会,结合 CLIR 的调查结果设计一个可扩展的原型系统。

关联数据为图书馆、高校和学术项目提供更强的跨界搜索和发现数字信息的能力。CLIR 的这项调研将为参加 2011 年夏斯坦福大学主办研讨会的代表提供背景信息。研讨会上,与会代表将研究关联数据的标准规范和实际需求,并在技术层面设计一个跨国、跨机构的原型系统以证明关联数据可以改善资源发现和资源导航的效果。研讨会之后,CLIR 将会发布调研报告,研讨会上产生的相关研究成果也将在网上公布。

"这对 CLIR/DLF 来说是一次重要的资助,既是资助者对于我们在问题研究和分析方面历来坚持的严谨态度的一种认可,也表明这次调研代表着一种新的研究和实践方向。"CLIR 主席 Chuck Henry 表示,"关联数据在组织和统一管理大量不同机构数字资源方面很有潜力,因此,CLIR 也一直将关联数据的大型解决方案放在战略核心位置。"

"我们正处在一个需要主流图书馆都认真考虑这项技术的关键时刻。"斯坦福大学图书馆员 Mike Keller 评论,"借助于 CLIR 的调研成果,我们希望这次研讨会之后能够为合作、分布式地研制元数据转换工具、元数据利用工具和可视化工具制定出具体的、可行的计划。"

(编译自:http://www.clir.org/news/pressrelease/11mellonpr2.html)

(本刊讯)