基于PROV模型的数据溯源

倪　静　著
孟宪学　审

中国石化出版社
HTTP://WWW.SINOPEC-PRESS.COM

基于 PROV 数据溯源

倪　静　著
孟宪学　审

中国石化出版社

内 容 提 要

本书在充分调研和总结现有研究工作的基础上，针对数据溯源模型在未来语义Web中的应用问题，提出一套语义标注和查询追溯实现方法，并加以验证；针对现有Web应用中溯源信息缺乏的问题，根据W3C推荐的溯源本体，结合已有的文本处理技术和语义服务平台，提出一套追溯方案，并通过实例进行验证评价；针对关联数据发布中溯源信息不足的问题，提出关联数据溯源的解决方案，并构建实验原型系统，检验合理性。本书深入浅出，逻辑清晰，具有一定的参考价值。

本书适合语义网和数据溯源方面的研究人员和学者使用；也可供相关专业院校师生参考使用。

图书在版编目(CIP)数据

基于PROV模型的数据溯源 / 倪静著. —北京：
中国石化出版社，2017.8
ISBN 978-7-5114-4571-1

Ⅰ.①基… Ⅱ.①倪… Ⅲ.①数据管理
Ⅳ.①TP274

中国版本图书馆CIP数据核字(2017)第171429号

中国石化出版社出版发行
地址：北京市朝阳区吉市口路9号
邮编：100020　电话：(010)59964500
发行部电话：(010)59964526
http://www.sinopec-press.com
E-mail：press@sinopec.com
北京富泰印刷有限责任公司印刷
*
700×1000毫米 16开本 8.75印张 151千字
2017年8月第1版　2017年8月第1次印刷
定价：48.00元

前　言

当前的网络环境正在向语义 Web 的美好愿景不断扩展和演进。然而，同时伴随数据源的不断增长，信息流动的不断加快，数据在传递过程中的频繁复制、演化，这些都给信息的可靠性、真实性和可信度带来了巨大的挑战。被 W3C 称之为语义 Web 最佳实践的关联数据发布越来越呈现如下特点：数量增长快，质量参差不齐，分布式发布使链接的复用越来越普遍，动态更新进一步导致数据的不一致。其根本原因在于起源信息的缺失。因此，在语义 Web 环境下，如何应用统一的数据溯源模型定位和查询起源信息，如何在相似网页中辨别真伪，以及如何在关联数据发布中增加溯源元数据，成为亟待解决的问题。

本书从目前语义 Web 研究中存在的瓶颈入手，以语义 Web、数据溯源、关联数据为理论基础，以语义 Web 应用为研究目标，综合运用文献研究、调查研究、系统分析、比较研究、归纳推理和软件工程等研究方法，在以下几个方面进行了相关研究：

(1) 开展了语义 Web 环境下数据溯源模型的比较研究。讨论了 DCMI 术语、OPM-O、PV、VoIDP、PROV-O 等目前国外主要的数据溯源描述语言，从来源和目的、资源描述角度、主要服务对象和解决的问题、标注方式、词表结构等方面分别对这些数据溯源描述语言进行了比较分析。

(2) 构建了 PROV 模型的应用情境。在分析 W3C 数据溯源推荐标准 PROV 的基础上，阐释该模型的主要功能，深入解析该标准的核心要素，构建了 PROV 模型的 Web 应用情境并进行描述；总结出该模型

的 Web 应用特征，即可解析性、语义性和可追溯性。这些工作对于推进国内同行在该标准的框架下，进行分布式环境下信息追溯和起源记录互操作的进一步研究具有一定参考价值。

(3) 验证了 PROV 本体描述的起源记录在未来 Web 应用中的语义表达和查询服务问题。在深入解析 Web 应用中起源记录的定位、传递模式、实现途径和实现模式的基础上，归纳总结出 Web 应用中起源元数据的四类定位发现机制与两类查询机制。结合语义标注 Web 页面和溯源信息表达技术，采用在线论文追溯案例，实现了包含 RDFa 语义标注起源记录的 HTML 页面显示，并通过可视化方式揭示起源，最后对案例中的起源记录查询服务问题进行了探讨。

(4) 提出了一种借助 PROV 本体自动化发现相似网页起源关系的实现方法。针对目前大多数已有的网页内容缺乏起源标注的问题，通过文档的变化过程分析，将文档定义为实体，并由多个语义属性进行特征描述，采用基于语义相似性的聚类方法发现文档间的变化关系，借助 PROV 本体发现文档的特征变化和责任人。通过属性识别抽取命名实体，建立文档属性的关系，链接到 LOD 云，借助通用语义本体发现起源的变化细节。最后以“转基因”为主题的网页为例，进行了相似网页内容检测的溯源验证。

(5) 构建了关联数据溯源的解决方案。从目前关联数据发布特征入手，提出数据溯源的必要性和着眼点。通过实体选择、粒度选择、工具选择的研究，建立了发布框架。基于 D2R Server 自动构建了用于教学和科研的关联数据，定制了溯源元数据，并实现了系统验证和 Sparql 终端查询。支持用户对本领域信息资源的共享、有效挖掘和及时跟踪。

数据溯源在我国尚属刚刚兴起的研究，基于 PROV 模型的数据溯

源研究荟萃集成国内外相关研究结论并进行了多方位的进一步探索，为解决数据真实性、可靠性及信任问题提供了理论与方法层面解决方案的学术参考。期待本书能引发读者对数据溯源等问题的深入思考，为进一步研究社交网络、大数据环境下数据溯源和数据质量问题奠定基础。

本书由倪静撰稿，孟宪学研究员审稿。在撰写过程中得到了许多专家学者的帮助指导，使本书涉及的研究工作都能够顺利开展并最终完成，这里一并表示感谢！同时感谢北京石油化工学院领导和同事在撰写过程中给予的支持和鼓励。感谢教育部人文社会科学研究项目“社交网络舆情演化的数据溯源及信任机制研究”（项目标准号：15YJAZH052）的资助。

本书内容属于探索性研究，不成熟之处在所难免，相关研究结果尚需在未来的实践中进一步验证和转化，恳请学术同行及各位读者批评指正。

目　录

1 绪 论

1.1 Web 向语义的 Web 演进及数据溯源

随着计算机技术和网络技术的发展，Web 已经成为全球信息的数据库，在 Web 向语义 Web 不断发展和演进的过程中，信息发布和传递呈现出新的特点：

第一，信息发布的频率发生了变化。信息的频繁发布和发布者对于关注度的要求，使得数据被不断复制和转发。随着数据流动的加快，原始信息的来源越来越难以捕捉，需要在新网页的发布中采用数据溯源标准和规范加以控制，从而定位和发现原始信息。

第二，信息发布的环境发生了变化。网络已成为一个开放的平台，不同质量的各种开放源均可提供信息。在传递的过程中，数据极易被修改或丢失。加之数据本身的不断演化，加大了数据验证的难度，数据的质量难以控制。在此背景下，对于已有的网络内容需要一套自动发现和追溯的方案，辨别真伪，获得信任。

第三，信息发布的数据架构发生了变化。语义 Web 技术推进了关联数据的发展，形成了多来源集成的全球数据网络，并已经被开发为联结不同数据源的多种数据架构，为用户提供所需要的信息。但是由于这些数据往往在分布式网络环境中更新，数据来源信息日益成为能否为科技人员提供可靠和信任服务的关键问题。

第四，信息发布的主体发生了变化。以最大的、跨领域的关联数据集 DBpedia 为例。DBpedia 旨在将 Wikipedia 的内容转化为结构化的知识，其目标是建立能够为语义网技术所应用的数据集。但是由于 Wikipedia 能够被任何一个人来编辑，辨别这些数据的来源与认定质量异常困难，其可靠性不免成为令人担忧的问题。而将关联数据应用于教学和科研时，这一问题显得尤为重要。

目前，溯源技术已广泛用于档案、艺术、考古学、医学、天文学、物理学等领域，从食品追溯(郑火国，2012)到网络环境下的数据新闻(Gray et al., 2013)、社会媒体(Barbier et al., 2013)、e-science 科学再现(Simmhan et al., 2005)和要求问责透明的财富追溯(Fax et al., 2013)等多领域的溯源应用正在兴起(倪静，2014)。也是数据溯源正在成为国外语义 Web 领域研究的热点之一。

万维网的发明者、W3C 主席 Berners-Lee 早在 1997 年就设想在用户界面有一个与文档相关的“oh-yeah?”按钮，当用户对所阅文档失去信任时点击按钮，通过软件直接或间接追溯元信息(meta-Information)，指出信任的理由。为了得到这样的结果，需要表达数据起源，并在 Web 环境下访问(Berners-Lee T., 1997)。

Berners-Lee 于 2000 年曾提出语义 Web 的体系结构(Berners-Lee T., 2000)，其主要的层次结构如图 1.1 所示。其中 Unicode 可允许任何人采用任何语言使用计算机的编码标准，而 URIs 用于唯一识别网上的概念，是 Web 架构中的资源识别机制。XML 层负责从语法上表示数据的内容和结构，通过使用标准的语言将网络信息的表现形式、数据结构和内容分离。RDF 解决的是如何采用 XML 标准语法无二义性地描述资源对象的问题，使得所描述的资源的元数据信息成为机器可理解的信息。Rdf Schema 使用一种机器可以理解的体系来定义描述资源的词汇，其目的是提供词汇嵌入的机制或框架，在该框架下多种词汇可以集成在一起实现对 Web 资源的描述。本体能够具体指明事物和彼此之间的关系。逻辑层从发布于 Web 的断言中生成新的知识。逻辑层一旦建立，便可以通过逻辑推理对资源、资源之间的关系以及推理结果进行验证，证明其有效性。证明层是跟踪逻辑推理的结果。加密技术特别是数字签名可用于 Web 上的真实性和非否认性认证。信任则可以用诸如此类的证明和加密技术建立。

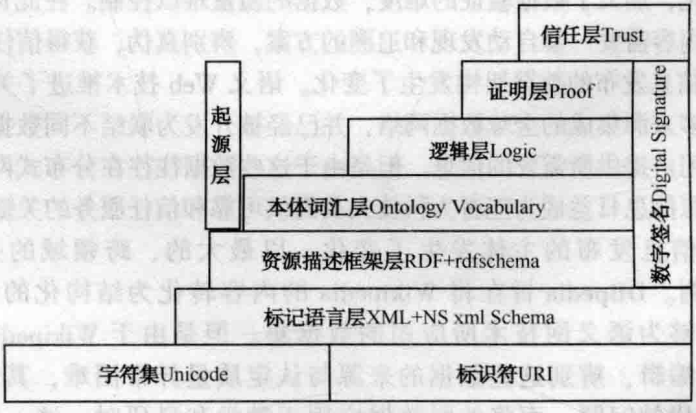

图 1.1　起源层在语义 Web 层次架构中的位置

自该架构提出以来，低层内容受到了大量的关注——RDF、OWL(即 Web 本体语言)、SPARQL，URIs 等，但是对于证明层和信任层的研究深度远远不够。特别是随着上述技术的广泛采用，大量结构良好、正确链接的关联数据出现的时候，会导致起源表达不正确或错误数据的情况。如何建立机制避免或减轻“不好的”关联数据的负面影响呢?

这需要广泛认可的数据溯源模型，要求该模型可用网络本体语言表示，遵从溯源的本体断言，可以在 Web 上表示和发布，并定义推理规则，允许独立于领域并对溯源的断言进行推理。通过对数据的来源表示和推理，判定数据的可靠性和可信性。由此可见，起源层在语义 Web 七层体系中，跨越多层：本体词汇、逻辑层和证明层，成为支持 Web 信任的主要因素。

数据溯源(也称数据起源)是新兴的领域，主要研究数据的演变过程，包括数据变迁的来源以及数据的权威性，数据生命周期的进化过程以及导致这些变化的因素。该技术已成为国外近年来语义 Web 领域的重要研究内容，举办了多个系列的数据溯源国际会议，语义 Web 环境下数据溯源研究的主题已由关联数据向大数据延伸。

有效的模型是数据溯源技术的关键所在，根据模型可以初步确定溯源的步骤，以及溯源的基本思路。建立于不同领域、不同行业的数据溯源模型曾经解决了数据库、工作流等方面的溯源问题。在语义 Web 技术研究中，W3C 的 PROV 一系列标准(Groth et al., 2013)推出以来，国际同行对 PROV 模型应用的高度重视，国外在 PROV 模型的领域扩展应用和在线验证方面已开展研究(Missier, 2013; Moreau, 2014)，这也是本书跟踪前沿、着力研究的重点。

根据中国互联网络信息中心(CNNIC)发布第 39 次《中国互联网络发展状况统计报告》，截至 2016 年 12 月，中国网民规模达 7.31 亿，互联网普及率为 53.2%。在网络越来越开放的情况下，一方面作为任何人都可以发布数据，另一方面，我国在数据溯源方面的研究相比于国外研究尚存很大差距，特别是在语义 Web 环境下的数据溯源技术研究更显远远不足。

数据质量是网络信息管理的基础，也是各项研究的生命线。越来越多的情况表明，不能明确数据产生、处理和演化的过程，信息的可靠性和真实性就会受到怀疑，往往得出错误和虚假的结论，甚至导致决策的失败。语义 Web 提出了 Web 自动化应用方面(如推理，聚合，或代理)的数据溯源问题，无论是使用混搭解决关联数据云中的问题，还是确定社会网络的信任问题，均需要找到源头信息。可见，起源是 Web 信息发布与消费的重要方面，也是语义 Web 向高质量和可信性发展的关键问题，迫切需要深入开展数据溯源及其模型在语义 Web 环境下的研究。

1.2 国内外研究进展

1.2.1 国外研究进展

数据溯源是一个新兴的研究领域，国外已经有很多大学和研究机构将数

据溯源作为研究课题，引起很多专家学者的高度关注。其研究历史可追溯到1986年(Becker R. A., 1986)，主要描述审计技术用于协助分析师理解和验证数据结果。最初，溯源系统是数据库领域的研究焦点，其研究始于1990年(Wang, 1990)，目的是发现对应于底层数据表的变化，实例化视图怎样及什么时候应该更新(Cui et al., 2000)。由于是明确定义的关系模型，它已被证明不仅能够从查询推导出精确的起源信息(Buneman et al., 2001)，而且可开发简明表达的表示形式(Green et al., 2007)。这些已经在系统中进一步扩展，例如斯坦福大学的Trio(Widom et al., 2006)，允许以纳入相关的不确定性，并可以通过起源进行多个记录的查询和传播。

数据溯源追溯项目和系统代表性的项目总结如下：操作系统层次上，以哈佛大学的PASS(provenance-aware storage system)(Muniswamy-Reddy, 2009)和SPADE(Support for Provenance Auditing in Distributed Environments)(Gehani, 2012)为代表，主要是通过检测应用软件的事件发生(例如处理的创建或I/O)，推断不同的数据之间的依赖关系；工作流层次上，VisTrails(Scheidegger, 2008)和ZOOM(Biton, 2007)是典型的工作流管理系统，能够追踪各种工作流执行的起源和自己(以VisTrails为例)的工作流程的演变，并且是工作流和可视化系统的结合；应用层次上，代表性系统Burrito(Guo, 2012)追踪用户-空间事件，同时也支持更多用户提供的标注。伊利诺伊香槟分校的SPROV系统(Secure Provenance)(Hasan, 2009)聚焦于起源的安全性，以平台无关的方式，实现了利用加密技术和存储技术的插件，可加载到现有的起源框架中去；在大数据层面，Lipstick(Amsterdamer, 2011)和RAMP(Park, 2011)都研究了大数据情境下追溯起源的问题。

国际上针对语义网和关联数据的组织和会议主要有：International Provenance and Annotation Workshop Series(IPAW)、Work-shop on the Theory and Practice of Provenance(TaPP)、eSI Workshop Developing a Set of Provenance Principles for Linked Open Data(PrOPr)、International Workshop on the Role of Semantic Web in Provenance Management(SW-PM)等。另外，W3C于2009年9月专门开设了W3C Provenance Incubator Group来研究语义网环境下的数据起源，其目标是“为语义技术、语义开发、语义标准的数据起源研究提供最新技术和发展规则”。

在Web和语义Web的研究领域可以分为以下几类研究：起源的网络发布、采用语义Web技术进行数据溯源、基于RDF的起源和溯源模型

研究。

(1)起源的网络发布

众所周知，网上信息揭示的基本原理是：①采用 URIs(Uniform Resource Identifers)识别全球资源；②采用协议(如 HTTP)访问资源。揭示起源的主要方法是超文本生成，RDF 视图和 Webdav 协议[1]。

Zhao(2003，2004)描述了从起源文档、数据、服务和工作流中动态生成的超文本文件。其目的是支持 Hendler 在《Science》中提出的科学网的愿景(Hendle J.，2003)，该网页通过本体推理、标注处理和链接插入而动态创建。Zhao(2004)等区别了起源的两个主要组成部分：附加到对(结构化、半结构化和自由文本形式)的标注和生成路径(工作流、查询或程序)。

SAM(Myers, 2003)是一个科学标注中间件，率先运用新兴的语义 Web 技术将初始捕获和存储的数据及元数据与表达进行分离，支持创建、使用元数据以及对元数据标注。特别是 SAM 还提供了捕获起源科学实验的电子笔记本，通过采用 Webdav 方法和 URI 标识符，实现了起源信息导航。采用起源浏览器生成起源浏览及 portlet 组件实现了图形可视化。

James(2007)和 Kwok(2006)提供了起源资源管理器，通过采用用户输入、语义推理和访问策略的结合，系统能够自动生成起源关系的个性化视图。起源信息主要由生成起源的系统(例如 Kepler[2] 和 Taverna[3])抽取，通过推论而创建用户视图。

(2) 采用语义 Web 技术进行数据溯源

语义 Web 技术研究中一直倡导促进起源的获取、表示和推理。一方面，RDF 允许通过 URI 引用资源，它的三元组结构简化了图形表示；另一方面，采用关联的查询语言 SPARQL 轻松地表达其查询，并可采用 OWL 定义本体和推理操作。

Zhao(2004)等从 4 个不同层次透视信息：组织层次(包括谁运行了工作流)、处理(一种事件日志，捕获输入输出)、数据(捕获数据演化)和知识(所有前续的批注)。Zhao(2004，2008)等主张采用 RDF 表达起源信息，采用 LSIDs(生命科学标识符)识别资源，利用本体提供此类数据语义的公共视图。这种数据网(Web of

[1] http://www.webdav.org/specs/rfc4918.html

[2] https://kepler- project.org/users/add_on_modules/provenance

[3] http://www. taverna.org.uk/documentation/taverna- 2- x/provenance/

data)可以用语义网可视化工具表示或通过浏览器浏览。

Chen(2006)等人使用术语“augmented provenance”来表示一条数据的起源和与导致此数据生成的处理有关的语义元数据。并对于如何从工作流构建环境和工作流执行引擎捕获此类语义元数据进行了解释。

正如 Zhao(2003)倡导语义服务，将领域知识融入表达，Sahoo 等(2008)使用术语“semantic provenance”表示溯源，其中纳入了领域知识和本体的支持，该方法已经被一些系统采用，数据比较依赖于具体的应用领域的专业数据。

Myers 等(2008)发现科学实验中数据与处理脱节，科学家必须手动采用异构的工具而不能集成，妨碍了实验再现的难度，在结果发布之前合作的往来记录(笔记、讨论和电子邮件)往往失去，大多数的原始处理和数据的追踪读者均不能访问。鉴于此，提倡使用语义内容管理系统，其中 Tupelo 是核心。Tupelo(Myers, 2008)，(Wang S., 2008)是一个提供 Web 访问协议和 Java API 的中间件，同时提供了与 OPM 模型的 RDF 映射接口。Wang 等(2008)开发了基于 OPM 模型的 GIS 专业应用软件。

Golbeck(2008)等展示了语义 Web 技术的灵活性来实现起源挑战。这一起源本体及相应的执行过程用 OWL 语言描述，用 SPARQL 实施查询。Miles(2006)利用 OWL 推理能力，确定实验的语义有效性。Christian 等(2006)提出了一个语义 Web 门户，采用提交者姓名和电子邮件等标注形式对数字图像进行批注并跟踪其起源。可浏览起源，以丰富用户的浏览体验。

在化工实验方面，Frey 等(2006)设想了基于 RDF 的用语义描述的世界，其中源头的标注规则被用于执行跟踪所有的信息，包括数字和物理状态变化的起源。处理数据时，标注处理的描述，使其起源明确有效。

在病毒学上，Balis 等(2008)开发了查询翻译工具 QUaTRo，允许用户通过向导，用户自己熟悉的语言和非技术概念构建起源和数据库的查询。

Tom 等(2012)提出了一种通过语义相似度发现粗粒度溯源的方法，但并没有给出具体的聚类算法和详细的系统实现方案。

(3) 基于 RDF 的起源

尽管许多作者主张用语义 Web 技术表达和查询起源，Klyne G 等(2004)持不同观点，他主张用 RDF 中三元组(以及版本和签名等问题)起源的识别问题。建议用命名图作为实体代表三元组集，用相关的起源信息标注。语义网允许 RDF