Computer Science and Application 计算机科学与应用, 2015, 5, 141-150
Published Online May 2015 in Hans. http://www.hanspub.org/journal/csa
http://dx.doi.org/10.12677/csa.2015.55018

Hans汉斯

Data Provenance on Semantic Web Based on PROV

Yanpeng Zhao[1], Chaofan Dai[1], Jing Ni[2], Lingmei Kong[3]

[1]Science and Technology on Information System Engineering Laboratory, National University of Defense Technology, Changsha Hunan
[2]Information Management Department, Beijing Institute of Petrochemical Technology, Beijing
[3]78046 PLA Troops, Chengdu Sichuan
Email: zypnole@sina.com

Received: May 9th, 2015; accepted: May 23rd, 2015; published: May 27th, 2015

Abstract

It is a research hotspot about how to establish a general rule of exchanging and sharing provenance information in heterogeneous systems. In this paper, we introduced the concept and usage of PROV, and put forward an idea about how to set up data provenance conceptual model and describe provenance information in ontology language. Finally, we established a scenario to test and verify the practical applicability and discussed the future direction in this field.

Keywords

Data Provenance, Semantic Web, PROV

语义网环境下PROV数据溯源技术应用研究

赵彦鹏[1]，戴超凡[1]，倪　静[2]，孔令梅[3]

[1]国防科学技术大学信息系统工程重点实验室，湖南 长沙
[2]北京石油化工学院信息管理系，北京
[3]78046部队，四川 成都
Email: zypnole@sina.com

收稿日期：2015年5月9日；录用日期：2015年5月23日；发布日期：2015年5月27日

摘　要

如何在开放环境下建立通用的，可以在异构系统之间交换、共享溯源信息的规则，是目前数据溯源技术的研究热点。本文引入了W3C提出的PROV数据溯源机制，重点探讨了在语义网环境下如何建立数据溯源核心概念模型以及如何用本体语言对溯源信息进行语义描述，构建了数据溯源实例进行验证，并展望了开放环境下溯源技术的研究发展方向。

关键词

数据溯源，语义网，PROV

1. 引言

数据溯源是关于实体、活动以及相关参与角色等信息的记录，对错误定位、质量保证以及信用提供等具有非常重要的意义。正如“溯源”这个词来源于古法语的“to come from”一样，溯源可以简单的定义为事物的起源以及寻找起源的过程。溯源技术自提出以来，即在数据库和工作流等领域得到了迅速的发展。虽然不同的领域对于溯源的需求和关注点并不一致，但总的思想都是要通过溯源，在数据共享时解决数据的可信度、质量、版本信息等问题，为用户提供可靠的来源语义信息，从而更加信任其获得或者使用的资源[1]-[3]。

2. 语义网环境下的数据溯源及 PROV

近几年，Web 技术发展迅速，数据流动方式的改变使得人们不得不更加关注所获得资源的可信度等问题。Web 环境下的数据有以下四个特点，也是其产生溯源需求的直接原因。首先，由于在 Web 上数据的传递和复制极为容易，导致数据流动速度加快；其次，数据的质量难以控制，在数据传递过程中被修改或者丢失更加普遍，数据质量难以控制；此外，Web 上的数据本身更新频率快，导致数据验证难度增大；最后，在分布式网络环境中，不同数据驱动的应用都会集合和融合一些数据，融合后的数据真实性和有效性将会大大降低[4]。如果我们可以探寻一种机制来描述事物或者数据的“生命周期”，对其在网络上的活动进行“追踪”并实现不同系统之间的信息共享，以上问题就可以迎刃而解。

2001 年，伯纳斯·李提出了语义网的概念，设想了不仅可以理解人类语言，而且可以使人与电脑的交流与人与人之间交流一样轻松的表达机制。语义网核心理念是通过给 Web 上的文档添加可以被计算机所理解的语义，从而使整个 Web 网成为一个通用的信息交换媒介[5]。添加元数据的思路与溯源技术中在数据变化过程中添加“注释标签”的思路一致，考虑将二者结合以求实现网络环境下的溯源。

随着语义网研究的深入，已经不再局限于溯源在单个领域中的应用，而是考虑将溯源信息以形式化的方式表达，并实现不同系统之间的互操作[1]。目前，针对 Web 环境下的溯源模型与术语集很多，包括 open provenance model (OPM)、provenance vocabulary 以及 PROV 等。其中，W3C 发布的 PROV 得到了专家和相关技术人员的肯定，是目前为止网络环境下溯源技术最成功的模型，具有广阔的发展前景。

2009 年 9 月，W3C 开设了 W3C Provenance Incubator Group (PROV-XG)组织来研究语义网环境下的溯源，设立了“为语义技术、语义开发、语义标准的溯源研究提供最新技术和发展规则”的目标，通过举办“国际溯源与注释大会(IPAW)”、发起起源挑战赛(the Provenance Challenges)等与全世界科学家、技术人员共同提出了 PROV。W3C 之后相继发布了一系列规范，包括 PROV 本体、数据模型等，现已成

为语义网环境下非常成功的溯源理念与模型表示方法。

PROV 系列共包括 12 份文档，针对普通用户、高级用户以及开发者三类用户。PROV-OVERVIEW 是对其他文档的总结介绍，PROV-PRIMER 介绍了 PROV 的基本概念，PROV-DM 提出了 PROV 概念数据模型，定义了描述溯源的一般词汇，可以在不同系统之间交换，PROV-O 是本体，采用允许将数据模型映射为 RDF 的 OWL2 语言所写。此外，其他相关文件包括：PROV-N，定义了方便用户使用的溯源标注；PROV-CONSTRAINTS，定义了 PROV 数据模型的使用约束集合；PROV-XML，定义了 PROV 模型的 XML 模式；PROV-AQ，介绍了起源的定位和查询机制；PROV-DICTIONARY，提出了由关键实体对组成的一类特殊集合；PROV-DC，提出了将 PROV-O 与都柏林核心术语互相映射；以及与 PROV 逻辑关系和应用机制相关的 PROV-SEM 和 PROV-LINKS [6]。

这 12 份文档中，PROV-DM 是核心，重点在是实现异构系统的互操作，实现溯源信息在不同系统之间的传递。PROV-O 是对语义网的借鉴，实现了 PROV-DM 提出的概念模型。本文将重点介绍 PROV-DM 和 PROV-O，探寻语义网环境下溯源技术的特点与规律，并结合实例进行分析研究。

3. PROV-DM

PROV-DM 是溯源的一般概念数据模型，将溯源信息转化为模型并重点关注不同系统之间交换溯源信息，这点在数据库和工作流领域中都未曾实现过。异构系统可以将本地溯源信息输出到这个核心模型，并支持有需求的应用程序输入、处理或者继续共享这些信息。

PROV-DM 分为核心结构和扩展结构，前者是构成溯源信息的本质，后者则使得溯源信息的描述更加具体化。PROV-DM 由六个模块组成，主要包括：1) 实体和活动，以及创建、使用或者结束时间；2) 实体来源；3) 实体产生与活动发生时代理所承担的责任；4) bundle 的注释，用于支持溯源信息之起源的机制；5) 将指代同一事物的实体属性进行链接；6) 针对成员形成逻辑结构的集合[6]。

3.1. 核心结构

核心结构包括溯源技术通用的三部分：实体、活动以及代理。实体可以指具体的物体也可以指抽象的概念。一辆车，一个文件，一个想法都可以看作实体。活动则指作用于实体上的行为，比如实体的使用、创建、处理、传播等。活动一般指时间段内的动作，比如现实生活中的开车、数字生活中的文件编辑等活动均有明显的开始与结束时间，并非瞬间完成。代理作为模型的基本要素，主要针对溯源关注质量、可信度的特性，指的是实体、活动的存在所承担一定角色的客体。例如某软件在处理文档活动中检查语法的使用，就可以看作是代理。代理回归到本质也属于特殊的实体或者活动。

在核心结构中，构成要素为“三类七关系”，即“实体、活动、代理”三者之间具有七种关系，基础的是活动使用(use)和产生(generate)实体，并受到代理在不同方面的影响。所有关系均为二元关系，用过去被动式(use 除外)表示，表示的是过去发生的事情，箭头方向由过去指向未来，具体定义如图 1 所示。

实体、活动和代理之间的关系具体解释如表 1 所示，为了与后文程序保持一致，保留了原关系命名。

3.2. 扩展结构

核心结构只描述本质溯源关系，对于更详细、复杂的关系则借助定义补充规则，即 PROV-DM 扩展结构。

扩展结构主要针对以下三种情况定义：

1) 类别划分

类别划分可以直接对核心结构进行划分。最典型的就是代理可以分为人(Person)、组织(Organization)

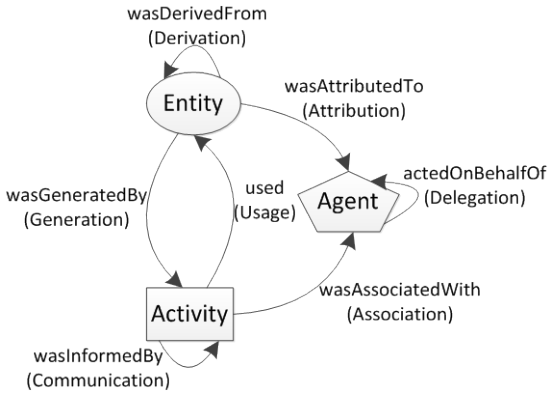

Figure 1. Entity, activity, agent and the properties among them
图 1. PROV-DM 中实体、活动以及代理之间的关系

Table 1. Properties of entity, activity and agent
表 1. 实体、活动以及代理之间的关系

关系名称	起点	终点	指代含义
Used	活动	实体	活动使用实体
Was Generated by	实体	活动	活动产生实体
Was Informed by	活动	活动	两个活动之间存在实体的交流
Was Derived by	实体	实体	活动使用的实体与产生的实体之间的关系，主要适用涉及活动未知或者不关注的情况
Was Attributed to	实体	代理	代理对实体承担责任
Was Associated with	活动	代理	代理在活动中承担责任
Acted on Behalf of	代理	代理	承担不同责任代理之间的关系

以及软件(Software Agent)三个子类，表示不同类型的代理。其他类似的还有“来自于(Derivation)”的子类版本(revision)，主要描述文件实体间的起源关系。类别划分是为了用 OWL2 语言描述 PROV 时更加直观与高效。

2) 关系扩充

核心结构中定义的均为二元关系，这种定义简单直观，但无法表示复杂的多元关系。以“来源于(Was Derived by)”关系为例，这是表示实体间比较高层次的关系，没有包含任何中间信息。用户对溯源的需求程度不同，就需要获得不同粒度的溯源结果，通过 PROV-DM 的扩展结构可以满足这个条件。具体做法是在二元关系中添加中间属性，用多个二元关系描述多元关系。

3) 扩展注释

主要针对具有时间、位置等细节信息的起源关系，通过运用资源标识符注释的方式更加具体的描述实体、活动等信息。扩展注释不仅可以获得更加细粒度的溯源信息，还可以更加准确得描述实体与属性之间的相关关系，提高溯源效率。

3.3. 具体应用

PROV-DM 实现了溯源信息模型化以及不同系统之间的交换。如何实现溯源信息在异构系统之间的交换是重点也是难点。具体做法是异构系统将本地溯源信息输出到该模型中，利用 RDF 的概念将具体内容用有向无环图进行表示与存储，之后想要使用该溯源信息的应用或者系统便可以进行信息的输入、处

理等活动。

下面，构建一个事例介绍 PROV-DM 的具体应用规则。背景为：某学生想写一篇研究报告，素材来源为某组织提供的表格数据，该学生通过某编辑软件对表格进行统计处理，并在得到的结果基础上完成了研究报告。在该实例中，“表格”、处理得到的“结果数据”以及“研究报告”均为实体，“处理”以及报告的“撰写”属于活动，而提供表格的“组织”、“编辑软件”及“学生”本身可看作代理。具体的关系为，处理使用表格，产生结果数据，撰写使用了结果数据，产生了研究报告，结果数据和表格之间以及报告与结果数据之间具有了“来源于”的关系，各自的代理分别为编辑软件与学生。由于编辑软件是学生使用的，所以学生与软件之间也具有间接的代理关系。如图 2 所示，可以用有向无环图展现。

4. PROV-O

PROV-O 即 PROV 本体(PROV Ontology)，定义了用 OWL2 网络本体语言来编码 PROV 数据模型(PROV-DM)的方法。PROV-O 与 PROV-DM 相对应，主要包括组成本体的类、属性以及用户权限限定集三部分。PROV 本体提供了不同领域实现溯源应用的基础，可以再现、交换、整合异构系统的溯源信息，构成溯源信息基于网络应用条件下交换的框架。由上文可知，PROV-DM 提出了一系列描述起源信息的概念，并建立了通用概念模型，而 PROV-O 在此基础上运用 OWL2 语言将该模型进行了本体化映射。

本体的概念类似于集合，即将一系列具有共同特性的归为一类。人与汽车就是典型的两类，结合 PROV-DM 概念来说，属于两类不同的实体。“拥有(hasOwner)”则是联接这两个实体的属性，且该属性具有方向性，由人指向汽车。OWL2 语言用不同的前缀表示不同的命名空间，每一个命名空间可以看作一个数据的集合。表 2 列举了在 PROV-O 中常用的前缀代表的命名空间。

同 PROV-DM 相对应，PROV-O 分为基础形式、扩展形式以及资格关系形式。可以根据需求选择。

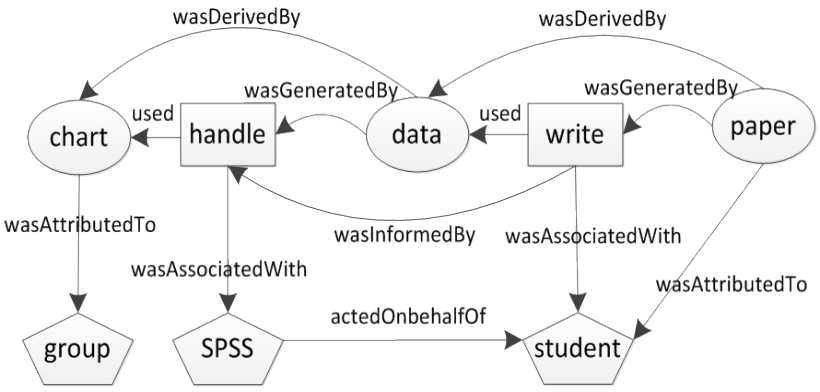

Figure 2. PROM-DM relations in directed acyclic graph
图 2. PROV-DM 核心结构的有向无环图

Table 2. Common prefix and namespace
表 2. PROV-O 中常用的命名空间及 OWL2 中的前缀[7]

前缀	命名空间(用资源标识符表示)	定义
xsd	http://www.w3.org/2000/10/XMLSchema#	XML 命名空间
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	RDF 命名空间
owl	http://www.w3.org/2002/07/owl#	OWL 命名空间
prov	http://www.w3.org/ns/prov#	PROV 命名空间

4.1. 基础形式

PROV-O 的基础形式对应 PROV-DM 中的核心结构，可以满足建立基础溯源描述的需求。OWL2 描述的类和属性采用 prov:classname 或者 prov:propertyname 的形式表示，如 prov:Entity、prov:used 等。运用 OWL2 规则表示活动 A 使用实体 E 则是如下格式：

@prefix prov: <http://www.w3.org/ns/prov#>.

E a prov:Entity; A a prov:Activity; prov:used E.

4.2. 扩展形式

扩展形式对应 PROV-DM 的扩展结构，可以更加详细地描述溯源信息。由于粒度的需求由用户需求决定，所以至今为止尚未形成一个标准的扩展规范。W3C 综合多个会议的成果以及现实需求，提出了一个“不完全”的扩展形式，可以实现绝大部分溯源粒度的需求。具体如表 3 所示。

对基础形式的各要素进行扩展，得到的主要是类的子类、属性的子属性以及超属性。代理的子类分为组织(prov:Organization)、人(prov:Person)以及软件代理(prov:SoftwareAgent)，实体的子类为束(prov:Bundle)、集合(prov:Collection)以及计划(prov:Plan)。此外，prov:wasDerivedFrom 的子属性对“起源于”进行了更加具体的描述，包括引用自(prov:wasQuotedFrom)、修订于(prov:wasRevisionOf)等。此外，还有实体的位置、代理的角色等也在扩展形式中进行了定义。

4.3. 限定性关系形式

由于 RDF 语言是用二元关系来表示网络上的信息资源的，这就导致在进行溯源描述时有的信息无法描述。比如，活动 A 产生实体 E，RDF 图的表示方式是 $E \xrightarrow{wasGeneratedBy} A$，这种表示只能提供“E 来源于 A”的信息，却无法描述这个产生活动的时间、该实体的其他情况等用户可能同样关注的信息。我们称这种情况为“非限定性(unqualified)”关系形式。PROV-O 定义了限定性(qualified)关系形式来解决。

图 3 表示了限定性关系的通用表示形式。假设资源 r2 与 r1 之间存在关系 prov:XXX。r1、r2 可以是实体、活动或者代理，prov:XXX 可以代表 used、wasGenerated 等关系。在 r1 和 r2 之间添加一个资源 x，使溯源图变为 r2->x 以及 x->r1，并将所需表示信息以注释的形式添加在 x 上，实现对多元关系的描述。

假设关系 XXX 代表 used，r2 为活动 a，r1 为实体 e，x 为使用关系 u，基础形式表达为：

:a a prov:Activity.

:e a prov:Entity； :a prov:used :e.

通过添加中间资源 x，“限定性”关系形式表达为：

:u a prov:Usage.

:a prov:qualifiedUsage :u.

:u prov:entity :e； :u prov:atTime ^^xsd:datTime.

x 不仅可以增添时间等标注，还可以作为实体或者活动等引申新的溯源关系。Luc Moreau 在《An introduction to PROV》一书中进行了详细的描述[8]。

总的来说，基础形式表示最基本的实体、活动以及代理之间的关系，扩展形式在此基础上增加了关系类别以及表示具体信息的注释，限定性关系形式则进一步对各关系的具体信息进行了注释。

4.4. PROV-DM 结合应用的设想

W3C 提出 PROV-O 主要是将 OWL2 语言强大的机器可读性能力应用到数据溯源信息描述的需求中。PROV-O 从本质上说与 PROV-DM 是两种不同的描述方式，但我认为可以将二者结合起来使用，从而更

Table 3. "Unfinished" extended terms by W3C
表 3. W3C 提出的"不完全"扩展形式

基础形式	扩展形式
代理	组织、人、软件
实体	束、集合、计划
无	位置
无	角色

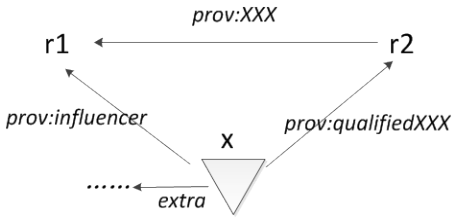

Figure 3. Qualified terms pattern of PROV-O
图 3. PROV-O 的限定性关系形式

好地实现对数据溯源信息的描述，提升数据溯源能力。

语义网的最终设想中，信息都被赋予了明确的含义，机器可以自动地处理和集成网上可用的信息，OWL2 是在语义网背景下的描述语言，侧重于描述 Web 文档中术语的明确含义和它们之间的关系。可以将 PROV-DM 与 PROV-O 结合起来顺序应用来进行溯源信息的表示与存储。PROV-DM 侧重于将信息用二元关系进行建模，用"三类七关系"以及其扩展关系全面而又简洁地描述溯源信息；PROV-O 则可以将概念模型用描述语言来表示，而且这种语言可以用于各种应用，并具有人与机器均可读的优点。下面通过构建具体案例来探讨如何将 PROV-DM 与 PROV-O 结合实现语义网环境下溯源信息的表示与存储。

5. PROV 实例应用

构建案例具体应用 PROV：某学生想撰写关于溯源技术的分析报告，原始数据主要来自两个专业组织。该学生借助分析软件对原始数据进行处理，完成了初稿。之后结合某网站发布的最新数据对文章进行了修改，完成终稿。由于上述过程涉及到不同的组织、机构以及处理过程，如果没有溯源的需求则可能最终仅有终稿的保留。数据溯源的目的则是使整个过程可以实现尽量大程度的追溯。通过上文提到的将 PROV-DM 与 PROV-O 结合使用的方法，首先用有向无环图的形式构建溯源数据模型(见图 4)，对整个过程进行描述，随后用 OWL2 语言进行编写，实现对整个过程的记录，并支持一定粒度的溯源查询或者定位。

接下来将上述概念模型用 PROV-O 进行本体化语义描述，目的是使溯源信息记录在可直观展现的基础上实现可编译，由于篇幅所限，只选取本体描述中具有代表性的进行编写与解释，原理类似的不再赘述。

1) 命名空间定义。

本实例用到的命名空间以及前缀代表主要包括：

@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .

@prefix owl: <http://www.w3.org/2002/07/owl#> .

@prefix prov:<http://www.w3.org/ns/prov#> .

@prefix :http://example.com/

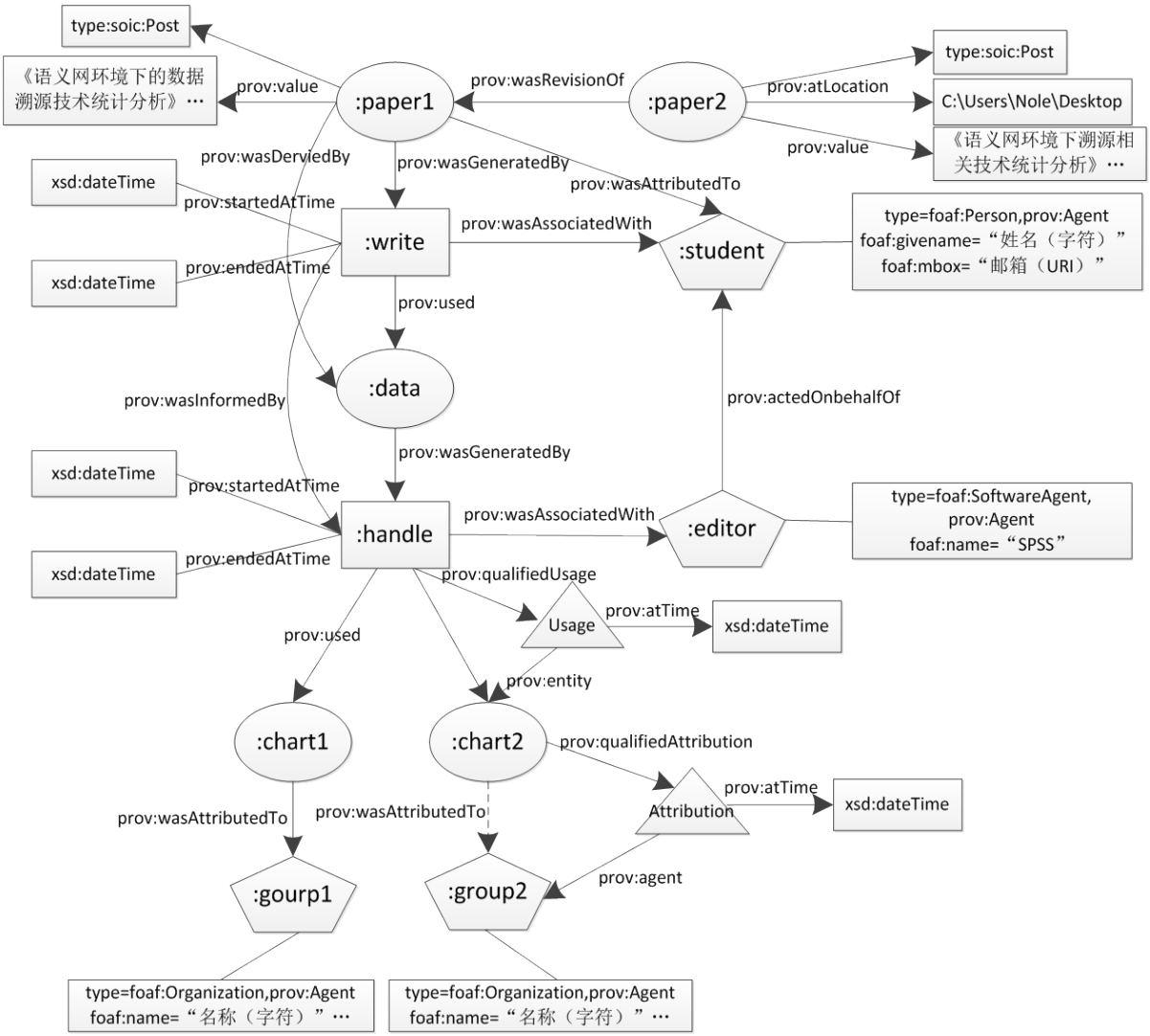

Figure 4. Conceptual data model of PROV-DM
图 4. PROV-DM 数据溯源概念模型

2) 实体、活动以及代理的定义方式，分别选取实体 chart1、代理 group1 以及活动 handle 为例：

```
:chart1
  a prov:Entity;
:group1
  a foaf:Organization,prov:Agent;##foaf 为一个 RDF 词汇表，表示朋友的朋友，在 prefix xsd 中定义。
        foaf:name "World Wide Web Consortium";
:handle
     a prov:Activity;
```

3) 实体、活动以及代理之间的属性关系，以活动 handle 与实体 chart1 以及 chart2 之间的关系为例：

```
:handle
     a prov:Activity;
  prov:used      :chart1;##非限定性关系形式
```

```
prov:qualifiedUsage[
a prov:Usage;   prov:entity      :chart2;
      prov:atTime "2013-12-09T13:02:00Z"^^xsd:dateTime;]##限定性关系形式
```

如第 4.3 以及 4.4 章所述，上文前三行采用的是非限定性关系描述形式，表示活动 handle 使用实体 chart1，后三行则采用限定性关系描述形式表示 handle 与 chart2 之间的关系，通过增加中间量 Usage 来为活动增添时间等属性。

4) 直接对实体、活动以及代理进行标注式描述，完善溯源信息。

```
:student
    a foaf:Person,prov:Agent;
      foaf:givenName "Nole";
      foaf:mbox <nole@gmail.com>
```

本例中代理 student 在定义时可以直接添加姓名、邮箱等具体标注，完善溯源信息。

```
:paper2
  a soic:Post,prov:Entity;
      soic:titile "语义网环境下数据溯源应用与研究";
 prov:value "溯源是关于实体、活动及相关参与角色…"
      prov:atLocation <C:\Users\Nole\Desktop>;
 prov:wasRevisonOf:   paper1;
:<C:\Users\Nole\Desktop>
 a prov:Location;
```

在定义 paper1 与 paper2 之间的关系时同样可以分别在定义其本身时添加标注，完善溯源信息。

6. 总结展望

本文主要介绍了如何结合语义网相关技术进行数据溯源，重点探究了 PROV 的核心模型和本体的具体原理，构建了实例来进行应用验证，完成通用溯源工具的构建与应用。

现如今关联数据、网络新媒体等新的信息环境使数据结构、数据流动方式都发生了重大转变，数据溯源技术的重要性越来越突出。结合语义网技术进行通用溯源技术研究是目前新兴的研究热点，PROV 可以作为描述和交换起源信息的通用工具。此外，PROV 为基础进行数据溯源还有许多其他可以深入研究的问题。比如，在 HTML 页面中表达溯源信息和实现可视化[8]，提供自动标注技术来实现扩展结构的相关注释[1]等。PROV 还在不断的完善之中，但作为通用溯源技术的一个重大突破，无疑为数据溯源技术打开了一个新的方向。

基金项目

本文系“十二五”装备预先研究项目(名称保密) (项目编号：513060403)研究成果之一。

参考文献 (References)

[1] 戴超凡 (2002) 数据仓库中数据志跟踪的理论与方法研究. 国防科学技术大学, 长沙.

[2] 明华, 张勇, 符小辉 (2012) 数据溯源技术综述. *小型微型计算机系统*, **9**, 2-7.

[3] 戴超凡, 王涛, 张鹏程 (2010) 数据起源技术发展研究综述. *计算机应用研究*, **9**, 2-6.

[4] 沈志宏, 张晓林 (2011) 语义网环境下数据溯源表达模型研究综述. *现代图书情报技术*, **4**, 1-8.

[5] Amit, S. and Cartic, R. (2003) Semantic technology in action: Ontology driven information systems for search, integra-

tion and analysis. *IEEE Data Engineering Bulletin*, **1**, 12.

[6] Luc, M. and Paul, G. (2012) Provenance: An introduction to PROV. Morgan & Claypool Publishers, Washington DC.

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

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