Research and Application of data provenance based on PROV

Yanpeng Zhao[1, a], Chaofan Dai[1,b] and Xiaoyu Zhang[1,c]

[1]National University of Defense Technology, Changsha, China

[a]871029507@qq.com, [b]cfdai@nudt.edu.cn, [c]zypnole@sina.com

Keywords: PROV, data provenance, ontology

Abstract. Focus on the exchange and sharing of data provenance information between heterogeneous systems, this paper introduced the concept and usage of PROV-DM and PROV-O, and put forward an idea about the combination of PROV-DM and PROV-O to describe provenance information, then established a scenario to test and verify the practical applicability. Finally, the paper discussed the future direction in this field.

Introduction

Data provenance is a record about people, institutions, entities, and activities involved in producing, influencing, or delivering a piece of data or a thing, which is very important for error locating, quality guarantee and credit. Provenance can be directly defined as the origin of thing and the process to find the origin. Technology of data provenance developed rapidly in fields of database and workflow since it has been put forward. Although the emphasis of provenance is different in different fields, the common idea is the same, which means solving problem of credit, quality and version information when sharing data and providing reliable provenance information for the users [1][2][3].

Data Provenance and PROV

As the development of Web technology, people began to pay more attention to that whether the resource they got can be trusted. There are four characters of data on Web environment, which are also the direct reasons of the need of data provenance. Firstly, because of the transfer and copy of data on Web becoming very convenient, the speed of data flow is very fast; Secondly, it is hard to control the data quality, which means the data will be easily modified or even lost in data transmission; Thirdly, the high update frequency of data on Web results in the highly difficulty of data verification; Lastly, in distributed network, different data drivers will conduct data aggregation or data fusion, and the authenticity and validity of result data will be significantly decreased. [4] Since that, if we can find out a mechanism to describe the “life cycle” of things or data, track the related activities on websites, and share the information between different systems, we can readily resolve the related problems above.

In September 2009, W3C set up Provenance Incubator Group (PROV-XG) to conduct the research of provenance on semantic web, and proposed the goal about “to offer the latest technology and development rules for the research of provenance about semantic technology, semantic exploit and semantic standard ”, and then published PROV by holding “International Provenance and Annotation Workshop” and “the Provenance Challenges” under the support of scientists and technology experts.

PROV consists of 12 documents facing to users, developers and advanced separately[7]. Among the 12 documents above, PROV-DM is core and it focuses on the interoperability between heterogeneous systems to realize the provenance information between different systems. PROV-O provides a set of classes, properties, and restrictions that can be used to represent and interchange provenance information generated in different systems and under different contexts. In this paper, we will introduce the basic conception of PROV-DM and PROV-O, and find out a way to combine them to express and store provenance information, and then introduce the idea by an example.

PROV-DM

PROV-DM is a generic conceptual data model for provenance that allows domain and application specific representations of provenance to be translated into such a data model and interchanged between systems, and this function has not been realized in fields of database and work flow. Thus, heterogeneous system can export native provenance information into the core model, and applications need to use the information can import it, process it, or reason over it.

There are two kinds of PROV-DM, core structure and extended structure. Core structure is essence of provenance information, and extended structure aims at enhancing and refining core structure with more expressive capabilities. PROV-DM consists of six components covering different facets of provenance:(1) entities and activities, and the time at which they were created, used or ended; (2) origin of the entities; (3) agents, and the responsibilities the agents beard when entities were generated and activities happening;(4) bundles, the mechanism to support provenance of provenance; (5) properties to link entities that refer to the same thing; (6) collections forming a logical structure.[6]

Core Structure. The core structure contains three parts which are the generic of provenance technology, entities, activities and agents. Entity can be concrete objects and abstract concept. A car, a computer file, even an idea can be taken as an entity. Activity means something that acts upon or with entities, like the usage, creation, processing or transforming of entities; In general, activities means the action over a period of time, like driving car, or editing digital file all have a duration, not instantaneous. As the basic part of data model, agent mainly aims at the characters of quality and reliability, and it is the object which bears some responsibility for activities or entities. For example, a kind of software which is in charge of checking the use of grammar can be taken as an agent. In its essence, agents belong to special entity of activity.

Core structure is formed of three classes and seven properties. Entity, activity and agent are classes and there are seven kinds of properties among them. The basic properties are usage and generation. All of the properties are binary relation, and expressed by past passive tense except used. The property is a description of what has happened, and the binary arrow points past to the future, shown in Fig. 1.

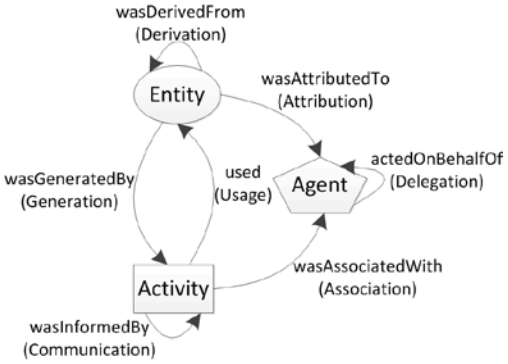

Fig. 1. Entity, activity, agent and the properties among them

The specific interpretation of the properties is listed in table1.

Table 1. Properties of entity, activity and agent

Name of property	Starting points	Ending points	Specific of property
Used	Activity	Entity	The beginning of utilizing an entity by an activity
WasGeneratedBy	Entity	Activity	Production of a new entity by an activity
WasInformedBy	Activity	Activity	Exchange of some entity by two activities, one using entity generated by the other one
WasDerivedBy	Entity	Entity	Relation between the entity that activity used and the entity that activity generated, transformation
WasAttributedTo	Entity	Agent	The ascribing of an entity to an agent
WasAssociatedWith	Activity	Agent	Responsibility to agent for an activity
ActedOnBehalfOf	Agent	Agent	The assignment of authority and responsibility to an agent to carry out a specific activity as a delegate or

representative

Extended Structure. The core structure only describe the essential relationship of provenance, while the extended structure focuses on the more specific and complex relationship and supports more advanced uses of provenance. There are three kinds of rules.

a. Subtyping

Subtyping can be applied to core types directly. The most typical example is that agent can be subtyped into three subclasses, Person, Organization and SoftwareAgent, which express different kinds of agents. Similarly, Revision is the subclass of Derivation, which means the resulting entity is a revised version of some original. In my opinion, subtyping can be used in PROV-O to make the description of provenance by OWL2 become more intuitive and efficient.

b. Extended Relations

All of the relations of core structure are binary relations, which are simple and comprehensive. But this kind of binary relation cannot express complex multiple relations. Take “wasDerivedBy” for example, this property is a high level relationship between two entities without any information about detailed provenance information. Actually, the demand for provenance is different for different users. By extending core structure, the PROV-DM can offer very detailed information for users. One method of the extending relations is adding intermediate property to binary relation, to describe multiple relations by several binary relations.

c. Extended Notes

It is mainly for the detailed provenance information about time, location and so on. By affixing extended resource identifiers or notes about specific information of entity and activity, the extended structure can not only express more fine-grained provenance information, but also can show the relationship between properties and classes to improve efficiency of tracing the provenance.

Specific Application. PROV-DM realized modeling and exchanging the provenance information, and it is very important and difficult to exchange provenance information between different systems with different expression mechanism. How to carry out it? Firstly heterogeneous system exports native provenance information into the data model, then expresses and stores the information by way of directed acyclic graph under the definition of RDF, and the applications can import the stored information and analysis it by its only mechanism.

Then we will set up an example to introduce how to use PROM-DM. A student wanted to write a research report, and the original material he used was some chart data. He handled the data by a kind of processing software, then finished his report based on the results he got. In this example, “chart”, “the result data he got” and “report paper” are entities, and “handle”, “write paper” are activities, and the “group who offer chart data”, the “software” and “student” can be regarded as agents. Then we can get the properties below: “handle” used “chart” and generated “data”, similarly, “write” used the “data” and generated “paper”. Since that there are derivation relation between “data” and “chart”, also between “paper” and “data”. And their corresponding agents are “software” and “student”. In addition, because the software is used by the student, there also is a relation of delegation between them. We can express these relations by directed acyclic graph, which is shown in Fig. 2.

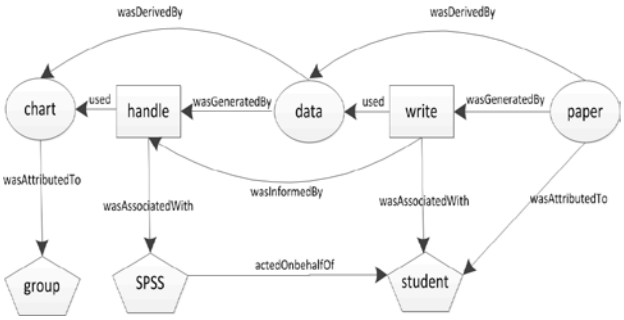

Fig. 2. PROM-DM Relations in Directed Acyclic Graph

PROV-O

PROV-O means PROV Ontology, which defines the OWL2 Web Ontology Language encoding of the PROV Data Model. Elements of PROV-O is corresponding to PROV-DM, mainly including the set of classes, properties, and restrictions that constitute the PROV Ontology. PROV Ontology provides the foundation to implement provenance applications in different domains, which can reproduce, exchange, and integrate provenance information generated in heterogeneous systems or under different contexts. We can know in chapter 2 that PROV-DM defines how to describe provenance information and sets up a generic conception model to represent provenance information in a variety of application domains. On that basis, PROV-O maps the model to ontology in OWL2 language.

The conception of ontology is similar to collection, both of them are putting a series of things with the same characters into one class. Person and Car are two different classes, and under the definition of PROV-DM, they are two different entities. “hasOwner” is the property which links them. In addition, the property is directed that from Car points to Person. The OWL2 uses different prefixes express different namespace, and each namespace can be regarded as a data collection. Common namespaces and their expression of prefixes shown in table 2.

Table 2 Common prefix and namespace

Prefix	Namespace (IRI)	Definition
xsd	http://www.w3.org/2000/10/XMLSchema#	XML Schema Namespace
rdf	http://www.w3.org/1999/02/22-rdf-syntax-ns#	The RDF Namespace
owl	http://www.w3.org/2002/07/owl#	The OWL Namespace
prov	http://www.w3.org/ns/prov#	The PROV Namespace

Being similar to PROV-DM, PROV-O can also be divided into different forms, the starting point terms, extended terms and qualified terms. Users can choose different forms according to their own needs.

Starting Point Terms. Like the core structure of PROV-DM, the starting point terms can satisfy the need the of basic provenance description. In OWL2, the class is described as prov:classname and the property is described as prov:propertyname, such as prov:Entity and prov:used. For example, activity A used entity E can be expressed as below under the rule of OWL2:

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

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

Extended Terms. The extended terms can describe provenance information more specific. Because the need of grained is decided by users, there haven’t been a unified standard about the extending level. Compounding many achievements of research and considering about practical needs, W3C put forward an “unfinished” extended terms which can satisfy most provenance needs in a relative simple method. The extended terms were shown in table 3.

Table 3. “unfinished” extended terms by W3C

Starting Point Terms	Extended Terms
Agent	Software, Person,
Entity	Bundle, Collection, Plan
None	Location
None	Role

Extended terms extend basic elements of the Starting Point terms with subclasses, subproperties, and a superproperty. To be Specific, subclasses of prov:Agent are prov:Organization, prov:Person and prov:SoftwareAgent. And prov:Bundle, prov:Collection and prov:Plan are subclasses of prov:Entity. In addition, the subproperties of prov:wasDerivedFrom is some detailed derivation description, including prov:wasRevisionOf, prov:wasQuotedFrom and so on. The location of entity, and the role of agent are also defined in the extended terms.

Qualified Terms. Both starting point terms and extended terms are binary relation, which sometimes are not sufficient to describe some situation. For example, we can express activity *A* generated entity *E* in RDF graph like $E \xrightarrow{wasGeneratedBy} A$, which can only offers the information about "E is generated by A", without any detailed information like the time at which the activity happened and the activity for which a delegation of responsibility took place. They are referred to unqualified terms. Then qualified terms were defined to offer further information in PROV-O.

In Fig.3, it illustrates the qualified terms pattern for an unqualified terms linking an influence *r2* to an influencer *r1*. Assume that there are a binary relation property prov:XXX between two resources *r2* and *r1*. *r1* and *r2* can be entity, activity or agent, and prov:XXX can express properties like used or wasGenerated and so on. The qualified terms pattern introduces a new resource *x* between *r1* and *r2*, which makes the binary relation become *r2->x* and *x->r1,* and then adds related information to *x* in pattern of notes. Since that, the information we cannot express by unqualified terms can be find on notes of x and the relation between *x* and the original resources *r1* and *r2*.

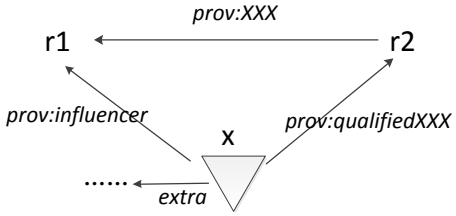

Fig. 3. Qualified terms pattern of PROV-O

Assume that *XXX* is on behalf of property used, then *r2* is activity a, *r1* is entity *e*, *x* is used *u*, so the unqualified terms pattern is:

:a a prov:Activity.

:e a prov:Entity； :a prov:used :e.

By adding intermediate resource *x*, we get the qualified terms pattern as below:

:u a prov:Usage.

:a prov:qualifiedUsage :u.

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

In addition, *x* can not only add notes about time, location and so on, but also can act as entity or activity for expounding new provenance relations.

In general, the starting point terms express the most basic relation among entities, activities and agents. On this basis, the extended terms divides different subclasses of the basic relation and adds some specific notes. At last, the qualified terms further completing the notes by adding intermediate resource.

Application Example of PROV

A student wanted to write a research paper about provenance technology, and the original material mainly came from two organizations. The material data was in chart pattern. Firstly he compiled the material chart by data analysis software, and finished first draft using the worked data. Afterwards, the student revised the draft paper according to some latest data coming from a professional website, and then finished the final draft. Because of the process being involved different organizations, groups and processes, only the final draft would be left if there was no requirement of provenance. Aim of provenance is realizing the tracing of the whole process as far as possible. By means of combining PROV-DM and PROV-O above, the provenance information can be described like this as below. Firstly set up a data model in directed-acyclic pattern to describe the whole process with classes and properties, and then compile the data model with OWL2 language. In Fig. 4, there are basic and extended descriptions, which can be chosen by users themselves. The whole process was described and stored in a standard way and compiled by OWL2, which is system independent and can be recognized and used by different systems and applications.

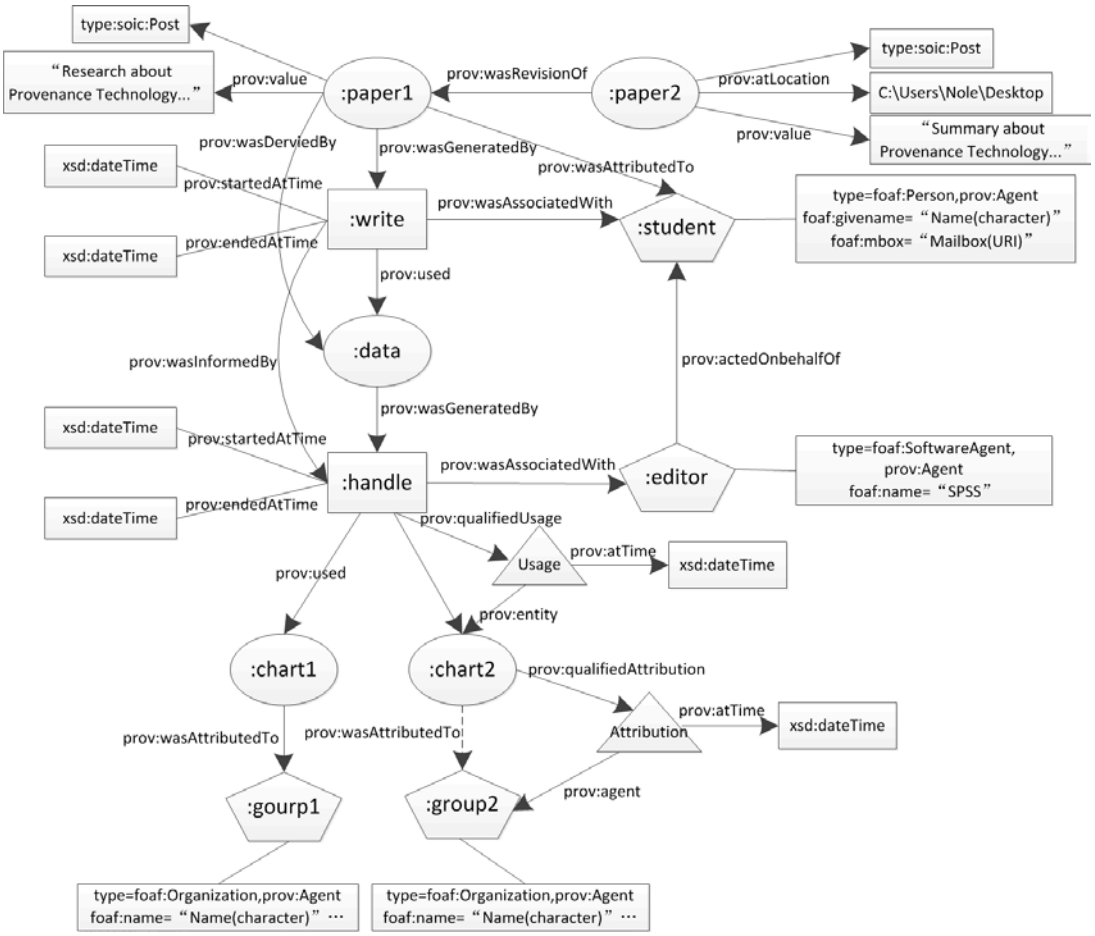

Fig. 4. Conceptual Data Model of PROV-DM

```
    @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#> .
:group2
     a foaf:Organization,prov:Agent;
     foaf:name "World Wide Web Consortium";
:chart2
     a prov:Entity;  prov:qualifiedAttribution[
    a prov:Attribution;
     prov:agent   :group2;
     prov:atTime "2014-02-14T00:00:00Z"^^xsd:dateTime;]
##describe the entity chart2, agent group2 and relation between them;  group1 and chart1are same, omit here
:handle
     a prov:Activity;
     prov:used    :chart1;##the relation of handle chart1(in unqualified terms pattern )
     prov:qualifiedUsage[
     a prov:Usage;  prov:entity    :chart2;
     prov:atTime "2013-12-09T13:02:00Z"^^xsd:dateTime;]
    ## the relation of chart2(in qualified terms pattern)
:editor
     a foaf:SoftwareAgent,prov:Agent;
     foaf:name "SPSS";## describe the SoftwareAgent
     prov:actedOnBehalfOf    :student;
:student
     a foaf:Person,prov:Agent;
     foaf:givenName "Nole";
     foaf:mbox <nole@gmail.com>## describe the Person agent
    :write
```

```
      a prov:Activity;
     prov:used :data;
      prov:wasInformedBy :handle;
      prov:startedAtTime "2014-12-22T01:01:01Z"^^xsd:dateTime;
      prov:endedAtTime "2014-12-24T02:02:02Z"^^xsd:dateTime;
     prov:wasAssociatedWith :studnet;//describe the processed result data
:paper2
      a soic:Post,prov:Entity;
      soic:titile "Summary of Provenance Technology…";
      prov:value " Provenance is information about entity, activity and the roles…"
      prov:atLocation <C:\Users\Nole\Desktop>;
      prov:wasRevisonOf:  paper1;
     :<C:\Users\Nole\Desktop>
      a prov:Location;
```

Summary

We introduced the basic technology of provenance in semantic web and the principle of PROV-DM and PROV-O, and then put forward an idea about the combination to describe provenance information. At last, we established a scenario to test and verify the practical applicability.

Nowadays, as the development of linked data, network new media and so on, the structure of data and the pattern of data flow have changed a lot, and the technology of data provenance becomes more and more important. It is a research hotspot that under the semantic web environment to exploit universal provenance technology. In this paper we discussed about how to utilize PROV as a universal tool for description and exchanging provenance information. In addition, provenance research based on PROV has many other issues to further study. Such as, how to describe and visualize provenance information on HTML page, and how to provide annotation of related extending structure by automatic annotation technology [1].Although the PROV is not completed and in constant development, it is a major breakthrough about provenance technology which provides a new way helping people to trust what they have got.

References

[1] Chaofan Dai, "Theories and Approach of Data Lineage Tracing in Data Warehouse Environment," National University of Defense Technology. Changsha, pp. 23-26, Apr. 2002.

[2] Hua Ming, Yong Zhang, Xiaohui Fu, " Survey of Data Provenance," Journal of Chinese Computer Systems, 9rd ed., vol. 33., 2012, pp.2-7.

[3] Chaofan Dai, Tao Wang, Pengcheng Zhang, " Survey of Data Provenance Technique," Application Research of Computers. 2010, pp. 2-6.

[4] Zhihong Shen, Xiaolin Zhang, " Data Provenance Model in Semantic Web Environment: An Overview," New Technology of Library and Information Service. 2011,pp. 1-8.

[5] Amit Sheth, Cartic Ramakrishnan, "Semantic Technology in Action: Ontology Driven Information Systems for Search, Integration and Analysis " IEEE Data Engineering Bulletin, 2003, pp.12.

[6] Luc Moreau, Paul Groth. Provenance-An Introduction to PROV. MORGAN&CLAYPOOL PUBLISHERS. [7] Demetrios G Sampson. "Ontologies and the Semantic Web for E-learning",Educational Technology & Society, 2004, pp.26-28.

[7] Jing Ni, Xianxue Ni, "Location and Query of Provenance Information in Web Applications", Library and Information Service, 2014,pp. 97-10.