

PTSLGA: A Provenance Tracking System for Linked Data Generating Application

Kumar Sharma

Department of Computer Science & Engineering, University of Kalyani, Kalyani, West Bengal, India
E-mail: kumar.asom@gmail.com

Ujjal Marjit

Centre for Information Resource Management, University of Kalyani, Kalyani, West Bengal, India
E-mail: sic@klyuniv.ac.in

Utpal Biswas

Department of Computer Science & Engineering, University of Kalyani, Kalyani, West Bengal, India
E-mail: utpal01in@yahoo.com

Abstract— Tracking provenance of RDF resources is an important task in Linked Data generating applications. It takes on a central function in gathering information as well as workflow. Various Linked Data generating applications have evolved for converting legacy data to RDF resources. These data belong to bibliographic, geographic, government, publications, and cross-domains. However, most of them do not support tracking data and workflow provenance for individual RDF resources. In such cases, it is required for those applications to track, store and disseminate provenance information describing their source data and involved operations. In this article, we introduce an approach for tracking provenance of RDF resources. Provenance information is tracked during the conversion process and it is stored into the triple store. Thereafter, this information is disseminated using provenance URIs. The proposed framework has been analyzed using Harvard Library Bibliographic Datasets. The evaluation has been made on datasets through converting legacy data into RDF and Linked Data with provenance. The outcome has been quiet promising in the sense that it enables data publishers to generate relevant provenance information while taking less time and efforts.

Index Terms — Provenance, Semantic Web, Linked Data, LOD

I. INTRODUCTION

The underlying rule of the Linking Open Data (LOD) project is to provide useful and related information on the web. The data published using a Linked Data approach are open in nature, represented by RDF. Altogether, these data based on Sir Tim Berners-Lee's Linked Data principles [1]. Users can publish open data, for which, there is no guaranty of quality and accuracy of the information. Even though, publishers have been publishing their data. It contributes to an enormous growth of data on the data hub. Information may arrive from diverse sources and users can publish any kind of data without any constraint. Sometimes the links for data items found to be out-dated [2]. Due to open nature of the published data, there are various ways of revealing trust and reliability of data on the web [3]. Therefore, it brings challenging problems for the

consumers to get the desired data. Subsequently all, consumer applications need information to appraise the quality of data on the web. In such instances, the publishers should provide trustworthiness and validity information at each entity and dataset level.

Many organizations and private sectors have already contributed their resource towards LOD cloud. Most of the resources are converted from legacy data systems using different Linked Data applications. These applications perform the transition procedure, assist publishing dataset on the information hub and automate the operation of generating provenance information. Only the degree of granularity differs with them. Provenance should also be tracked at the triple level. We also argue that capturing provenance information after the conversion process would take additional time and efforts and get a low-quality provenance. This is mainly because the process related information such as when an operation began, when it ended, its input & output parameters are not available easily after generating a resource. Hence, these applications should track, store the provenance information and associate it to the RDF resources during execution. Sometimes legacy data systems also contain the provenance metadata attached with them. Such information should also be captured and retained during transitioning process.

In this article, we present how the provenance of RDF resources is tracked and stored at the time of generating RDF resources. We also present how the provenance information is disseminated on the web. The structure of this paper is as follows. Section 2 describes some related work on this field. In Section 3, we briefly describe provenance representation and uses of widely used provenance models in the domain of Semantic Web. Section 4 explains the basic aspects of representing provenance along with conceptual terms and describes the provenance capturing architecture. In Section 5 we present the way of disseminating, consuming provenance on the web, Section 6 presents the experimental evaluation, and Section 7 concludes the work.

II. RELATED WORK

Linked Data generating applications need data and workflow provenance to be tracked during creation of the RDF resources in order to provide a quality provenance. The provenance needs to be managed, in the same way, as the RDF resources are available on the web. However, most of the current practices support provenance generation only at the dataset level using VoID vocabulary (Vocabulary of Interlinked Datasets) [4]. An extension to the VoID vocabulary, VoIDP, which captures the information regarding workflows and activities involved in creating an RDF dataset, is discussed in [5, 6]. A suitable provenance model that captures information about web-based and the creation of data has been elaborately discussed in [7]. One can produce provenance information at both dataset and resource level through this model. The “Named Graph” concept has also been applied to deal with the provenance information concerning the links between the data items from various sources [8].

Public Key Infrastructure (PKI) principles [9] have been used to express the trustworthiness of the dataset by using private and public keys. They suggested the use of third party trust center or Certification Authority (CA) which issues digital certificates to verify each linked datasets. These certificates have been employed to identify the validity and the quality of the datasets. A metadata component [10] for the Linked Data publishing tools such as Triplify, Pubby and D2R Server is useful for generating metadata of a huge number of RDF dataset. The component relies on the metadata provided by the publisher and allows automatic generation and publication of the provenance metadata at the dataset level. The method of automatic discovery of high-level provenance using semantic similarity has been illustrated in [11]. The methodology relies on the clustering algorithms and the semantic similarity. This approach provides provenance at the document level. In [12] authors have presented an approach on automatic generation of the metadata based on VoID vocabulary. The result produced is dataset information such as statistical data and information about linked datasets. In [13] authors have used the concept of tracking provenance information through use of Version Control System (VCS) such as Github. VCS has been mainly used in controlling and tracking source code to facilitate teamwork. Each collaborator performs some task on data or files system and commits their changes and contribution. They track such commits and perform mapping with the W3C PROV data model along with other metadata. They have provided RESTful web service to offer the provenance of such workflows. Users only need to provide URL (Github URL) that point to a Github repository. This process provides provenance at the document and file level.

A framework has been explained for converting cultural heritage data into RDF [14]. Legacy data were initially stored in spreadsheet files. The tool XLWrap has been used to translate spreadsheet data into arbitrary RDF

graphs through mapping information. The converted RDF data are made available on the web accessible via SPARQL end-points. It uses two ways to disseminate the provenance or meta-data of the dataset: using VoID description of the dataset where it is published in a URL [22]. Another way is by describing used terminologies in the form of published vocabularies. For example, in [14] authors used the “?Hvor” vocabulary to represent the address information of the buildings in the Yellow List. Converting raw government data into Linked Data based on the LGD (Linked Government Data) Publishing Pipeline has been presented in [15]. The raw data are available on various formats such as JSON, CSV and TSV. They used “CKAN Extension for Google Refine” for sharing data while tracking provenance of the RDF data. This project extracts the workflow operations that are associated with the RDF data, which is purely based on process-oriented provenance. A system called BibBase [16] transforms bibliographic data stored in BibTeX files into Linked Data. RDF data are stored into a triple store which is available to be queried using SPARQL. Data in the BibBase comes from the various sources and BibTeX files. The provenance information, for each entity, is recorded by capturing the source of each entity and each link that are encoded with the entity. This method is based on data-oriented provenance where they track only the source of the data items.

III. PROVENANCE REPRESENTATION

The Provenance representation is a model, which assists users to express provenance of their data. Various provenance models have been evolved such as Open Provenance Model (OPM), Provenance Vocabulary, W3C Provenance Model (PROV), and VoID. The VoID is widely used in the context of Linked Data. It is a vocabulary for describing RDF dataset. VoID only deals with describing metadata of the RDF dataset. VoID has been applied to describe General Metadata (following the terms from Dublin Core), Access Metadata (information about how to locate and access the RDF data), Structural Metadata (Internal schema and technical features of the dataset), and the Description of the Linked set (a set of RDF Links). General Metadata follows the terms from Dublin Core Metadata Element set. Such as, title and description (dcterms:title, dcterms:description), licensing information (dcterms:license), creator and publisher of the dataset (dcterms:creator, dcterms:publisher). Access metadata gives information about how the RDF resources are accessed over the web. For example, SPARQL endpoints, RDF data dumps, Root resources, URI lookup endpoints, and Open Search description documents are some of the ways to access RDF dataset and RDF resources. Structural metadata provides internal structure and technical features of the dataset. For instance, the information about vocabularies used, the total number of RDF statements, entities are some of the technical features of the dataset. Oftentimes RDF resources have links to other resources from outside datasets. Such link sets are described using void:Linkset and void:target

properties. Hence, VoID plays a pivotal role to distribute provenance only at the dataset level. A separate provenance model is required to represent provenance for each RDF resources in Linked Data. PROV provides core data model for representing provenance using concepts such as Entities, Activities, Agents, Roles, Time, Usage and Generations. These main concepts enlighten about how an entity came into existence, what processes and activities are involved in generating the entity, who was involved in performing the activities, what other data items were used and at what time. These key concepts take part in different aspects while performing various

activities. Here, the main activity is data conversion, which converts source data items into target data items. Users who are involved in this conversion process are agents, which in turn are associated with a particular agent, an organization. So, using all these concepts publisher can model provenance of a data item. Fig. 1 shows the detail concept about how a data item is related to an activity, agent and other data items. To represent provenance of RDF resources we use PROV data model. VoID vocabulary is used to describe the provenance of the RDF dataset.

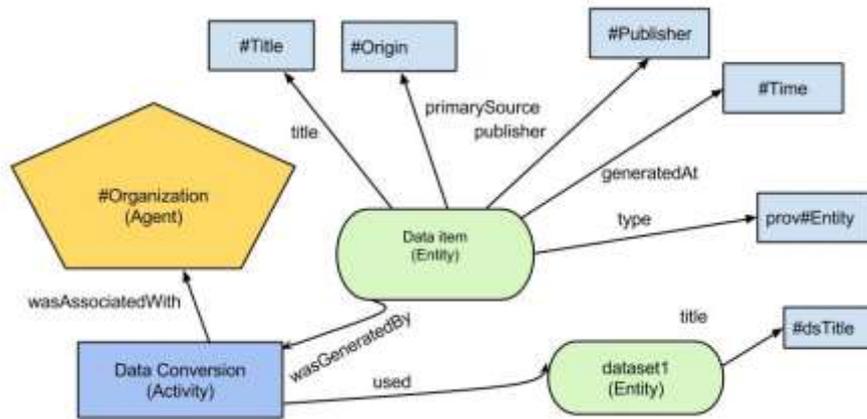


Fig. 1. Describing Provenance Information using Actor, Activity and other Entities in PROV.

IV. BASIC ASPECTS OF REPRESENTING PROVENANCE

RDF is a data-modeling framework for representing web resources. A web resource, identified by its URI is represented using different syntax and representation formats. To describe provenance information of RDF dataset and its resources a collection of source data and process related information are required. This information is available during generation of RDF triples. It would be difficult to obtain the process and operation related information at the later stage of the conversion process. Hence, the provenance information is recorded during generation of the RDF resources. In doing so, the information about processes, activities, agents and the operations are instantly available. In addition, the same process is responsible for generating provenance information. The proposed approach also records the provenance of the data items recorded by the previous system in the legacy data format. First we present below the basic concepts which are needed while representing provenance.

A. RDF Statement

The abstract structure consisting of Subject (S), Predicate (P) and Object (O) is called an RDF triple. Each such triples state that the Subject and Object are in some kind of relationship joined by the Predicate (P). Such statement in RDF is called an RDF statement. Let T denotes the RDF Triple, then

$$T = S \cup P \cup O \tag{1}$$

In RDF, the subject is always the resource that is being described. A resource can be of anything, a place, a person, and a book such that subject and predicate are always identified by URI whereas the object, which can be either a resource (identified by URI) or a literal value. In this work, we denote a subject by R.

B. RDF Dataset

RDF dataset is a collection of RDF triples or statements. It is also called a directed or labeled graph where subjects and objects are nodes and the predicate represents the arc. Let $\{T_1, T_2, \dots, T_n\}$ be a set of RDF triples. The RDF dataset D is defined as:

$$D = \{T_1, T_2, \dots, T_n\} \tag{2}$$

C. Provenance of Data Item

In generating data-items of any kind, the data item is associated with many things such as agents, activities, processes, and other used data-items. Hence, provenance of a data item, in the perspective of the agent, entity and process oriented provenance, is a collection of agents, activities, processes, and source data items. We represent the provenance for any data-item by P such that,

$$P = [A, V, F, S] \tag{3}$$

Where,

$A = \{A_1, A_2 \dots A_n\}$ is the set of all agents,

$V = \{V_1, V_2 \dots V_n\}$ is the set of all activities,

$F = \{F_1, F_2 \dots F_n\}$ is set of functions or processes that generates the data-item or relates a data-item to another data-item.

$S = [S_1, S_2 \dots S_n]$ is the collection of source of the data item, other used data items and the brief description about the origin of the data.

D. Provenance of RDF Resource

Provenance of RDF resource may be defined as tracking provenance of R where R is the resource or subject in an RDF triple. Provenance of R is defined as the combination of its agent, process, entity oriented provenance and the provenance generated by the previous data storage system. Such that,

$$P_R = P \bullet P_L \quad (4)$$

Where,

P is the agent, process, and entity oriented provenance.

P_L is the provenance information recorded by the previous system for legacy data.

E. Provenance of RDF Dataset

The provenance of an RDF dataset is a collection of general metadata, access metadata, structural metadata and the description of link set as well as the combination of its agent, process, and entity-oriented provenance, which is described by VoID. Such that,

$$P_D = [G, \alpha, \psi, \lambda] \bullet P \quad (5)$$

Where,

G is general metadata,

α is access metadata,

ψ is structural metadata

λ is the description of the link set, and

P is the agent, process, and entity oriented provenance.

The comprehensive architecture of the framework has been shown in Fig. 2. Based on the above concepts, for each data item, we capture the provenance information such as the source of each statement, associated activities, agents, the date-time information, the actors and the information about processes that were involved in the creation process. The source of the data provides information, which corresponds to the creational history of the data. The source of the data may not be available to the process and therefore the agent needs to enter it explicitly. The agent enters basic information such as the source of the input data, agent's address and the description about the input data and licensing information. Sometimes the provenance is also captured and recorded in the legacy data system. For example, in case of MARC 21 record, the Field 561 defines it as "Ownership and Custodial History" [20]. The provision for storing such information has been made and is combined with the agent, process, and entity oriented provenance. The agent is associated with the data creation and modification process. It is automatically created and stored into the store.

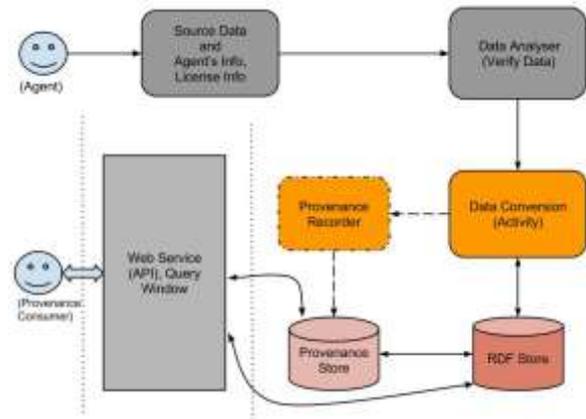


Fig. 2. Provenance Capturing Architecture.

Whatever data are being created they are associated with only one agent. The data conversion activity is the process, which creates data item and assigns relevant provenance information to it. Hence, the source data, date-time parameters, the agent, activities, and other entities are required to capture the provenance information. In the end, the conversion process collects all relevant metadata of the dataset and creates VoID file separately. By this, the Linked Data generation application becomes conscious towards provenance, making provenance tracking as a mandatory step for them. As shown in Fig. 2, the legacy data along with source information, agent's information as well as licensing information is entered by the user. In the next step, data analyzer analyses the data to be processed. This will verify whether the supplied data can be processed by the data conversion activity. The data conversion activity takes each data item, converts into RDF and Linked Data. As a whole, the provenance information is recorded by the provenance capturing method. A separate method is added which performs the job of provenance capturing. During the conversion process, for each data item, the corresponding provenance item is created and its certain attributes such as date-time, activities, and agents are assigned to the provenance data item. Any source data, which are linked to outside sources, such as `rdfs:seeAslo`, `owl:sameAs` links are treated as used data items. The provenance store communicates inwardly with the RDF store to retrieve external data sources that have been linked with the data. Each time the legacy data item is encountered, it is converted into RDF resource, integrated with other data sources from the web such as DBpedia, VIAF and then it is stored into the RDF store. Simultaneously, the provenance generator captures the provenance of the data item and stores into provenance store.

V. PROVENANCE PUBLICATION AND CONSUMPTION

Publishing provenance is a vital job for the data publishers. Making the availability of the provenance information on the web is always a concern. In some measure, publishers should provide references or access information of the location of provenance for each data

item. Provenance is valuable information that one must provide its access in order to ensure that data is trusted. Several choices have been made regarding provenance publication. Oftentimes VoID is used to express metadata of Linked Data. Using VoID, we can publish the provenance of the RDF dataset as a separate document on the web and then it is linked from the RDF document using `void:inDataset` property as discussed in [17]. The Provenance Access and Query (PAQ) working draft [18] have elaborated a number of possible ways for accessing provenance for individual data items. It is mentioned that the provenance information should be accessible, in the same way, as the resources are accessible on the web, by dereferencing the HTTP URI. It means that the provenance information is also a resource represented and described by RDF having dereference-able HTTP URI for each resource. The following are the two different ways for accessing the provenance:

A. Indirect Access

When provenance is not accessible using provenance-URI, or they are not accessible as the web resource, a Query Service can be used to serve the provenance information. Such query service, such as SPARQL service endpoint, processes the SPARQL queries submitted by the consumers. In such cases, the publishers should mention the provenance service URI in the RDF resource and HTTP response header field for the resource represented by RDF and HTML respectively.

B. Direct Access

Another way to access provenance information of the RDF resources is by dereferencing provenance-URI. Provenance-URI points to the actual provenance record generated and stored by the data publisher. Dereferencing this URI discovers the provenance information associated with the original RDF resource. Provenance URI has been embedded in the original RDF resource using “`prov:hasProvenance`” property for the resource, represented as RDF. It can also be associated with a resource by adding `<Link>` element followed by

`<prov#has_provenance>` in the HTTP response header field for the resource represented as HTML.

In the proposed work, we follow the direct approach. Provenance information for each resource has been defined using RDF following the PROV data model. An RDF resource represented by the URI `<BibliographicLinkedData/BibResources/37th_annual_meeting.>` has the resource description. In the resource, reference to the provenance URI has been added using property `<prov:hasProvenance>` followed by the provenance URI `<BibliographicLinkedData/Provenance/37th_annual_meeting.>`. Dereferencing the provenance URI results in the provenance description in RDF format considering the “Accept-Encoding” of the HTTP header field as “RDF/XML”. The provenance description tells that, the resource `<BibliographicLinkedData/BibResources/37th_annual_meeting.>` `prov:startedAtTime` “Sat Jul 05 23:29:45 GMT+05:30 2014”, `prov:wasGeneratedBy` the activity `</Provenance/ld_converter/>`, `prov:wasDerivedFrom` the dataset `</Provenance/ld_dataset/>`, it `prov:used` the resource “`http://viaf.org/viaf/147081590`” for generating the links and was `prov:endedAtTime` “Sat Jul 05 23:29:49 GMT+05:30 2014”. This resource `prov:hadPrimarySource` `<http://openmetadata.lib.harvard.edu/bibdata>` and has `DC:provenance` as “Boston, Mass.” from the previous data storage system which was published by “American Society of Landscape Architects”. The activity `</Provenance/ld_converter/>` whose title is “Legacy Data to Linked Data Converter” `prov:wasAssociatedWith` `</Provenance/ld_agent/>`, created by “University of Kalyani”. The agent `</Provenance/ld_agent/>` is type of “`prov#Person`” and was `prov:actedOnBehalfOf` another agent `</Provenance/CIRM/>` of type “`prov#Organization`”. In this way, provenance has been generated for each resources.

Table 1. Evaluation Results (Legacy Record to RDF with Provenance)

Dataset	No. of Legacy Record	Time to convert Legacy record to RDF without Provenance (in Seconds)	Time to convert Legacy record to RDF with Provenance (in Seconds)	Number of RDF Resource (in %)
Dataset 1	100000	28s	52s	96%
Dataset 2	100000	24s	50s	97%
Dataset 3	100000	21s	45s	87%
Dataset 4	100000	16s	39s	100%
Dataset 5	100000	18s	40s	94%
Dataset 6	100000	23s	48s	87%

VI. EXPERIMENTAL EVALUATION

Harvard Library Bibliographic Dataset has been used as legacy dataset, provided by Harvard Library [21] for the sake of experiment. Library aims at providing open

metadata in the library domain to support learning and research for the students as well the researchers. The datasets consist of bibliographic records in MARC 21 format, exported from Harvard's Library for public use. The datasets contain more than 12 million bibliographic

records of different categories such as books, journals, electronic resources, audios, videos and other materials. The proposed work is an extension of the previous work [19]. The method has been improved and it takes less time to process legacy resources than previous work. Six different bibliographic datasets have been processed for the sake of experiment. Each dataset is of different size,

from which a varied number of RDF resources and provenance information have been generated. As shown in the Evaluation Table 1, to convert first 100000 legacy records without having provenance and links to outside sources it took less than 30 seconds. However, it took around 50 seconds with provenance records but without having links.

Table 2. Evaluation Results (Legacy Record to Linked Data with Provenance)

Dataset	No. of Legacy Record	Time to convert Legacy record to Linked Data without Provenance (in Seconds)	Time to convert Legacy record to Linked Data with Provenance (in Seconds)	Number of RDF Resource (in %)
Dataset 1	100000	31243s	31904s	96%
Dataset 2	100000	27209s	29138s	97%
Dataset 3	100000	28971s	31278s	87%
Dataset 4	100000	30929s	39614s	100%
Dataset 5	100000	32138s	33860s	94%
Dataset 6	100000	19260s	21156s	87%

It has also been observed that the time for converting legacy records without having RDF links is always very less. This is because for every record the link generation method has to fetch links from outside sources. Currently, it fetches similar links from DBpedia and VIAF sources and subsequently attach the links to RDF resource using the RDFS:seeAlso property. Hence, in order to keep links for every record we need to query outside sources and wait until the response comes back. This is very time-

consuming task. As shown in Table 2 & Fig. 3, time needed to convert legacy data into Linked Data is always much larger than the time needed to convert into RDF without links.

The experiment has been performed on a 64-bit 2 GHz Intel Core i7 processor having 4GB of RAM running on Mac OS X 10.8.3. Jena 2.10 framework has been chosen to process RDF data along with Jena TDB to store RDF and provenance resources.

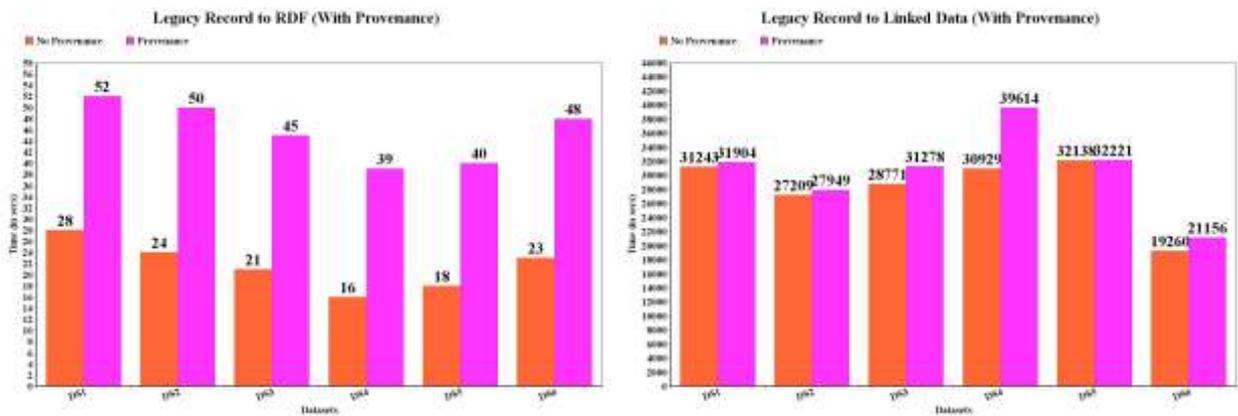


Fig. 3. Legacy Record to RDF and Linked Data conversion for first 100000 records.

VII. CONCLUSION

In this article, we have shown how the provenance information of the RDF resources can be generated during the conversion process. Since many Linked Data generating applications are not aware of provenance information of RDF resources. During the conversion of legacy data to Linked Data the provenance needs to be captured and recorded instantly. Many consumer applications need a technique to trace metadata or provenance of the data items on the web. For this we have followed the direct approach, mentioned in

Provenance Access and Query (PAQ) working draft [18], to disseminate the provenance information. We have shown how an RDF resource can be linked to its provenance-URI. Provenance-URI is dereference-able on the web and is accessible by both HTTP and RDF. We believe that by tracking provenance of RDF resources during the conversion process will help publishers in generating provenance while reducing the time and efforts. The future work includes the conversion of the legacy data from other data formats such as CSV keeping their provenance, versioning and change information.

REFERENCES

- [1] C. Bizer, T. Heath, and T. Berners-Lee. "Linked data-the story so far". *International journal on semantic web and information systems*, 2009; 5(3), 1-22.
- [2] A. Schultz, A. Matteini, R. Isele, P. N. Mendes, C. Bizer, and C. Becker. (2012). "LDIF - A Framework for Large-Scale Linked Data Integration". In 21st International World Wide Web Conference (WWW 2012), Developers Track, Lyon, France.
- [3] G. Ciobanu, & R. Home. "A provenance tracking model for data updates." arXiv preprint arXiv:1208.4634 (2012).
- [4] K. Alexander, and M. Hausenblas. "Describing linked datasets - on the design and usage of VoID, the Vocabulary of Interlinked Datasets". In *Linked Data on the Web Workshop (LDOW 09), in conjunction with 18th International World Wide Web Conference (WWW)*. 2009.
- [5] T. Omitola, L. Zuo, C. Gutteridge, I. C. Millard, H. Glaser, N. Gibbins, and N. Shadbolt. "Tracing the provenance of linked data using VoID". In *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 2011. ACM, p. 17.
- [6] T. Omitola, N. Gibbins, and N. Shadbolt. "Provenance in Linked Data Integration", 2010.
- [7] O. Hartig. "Provenance Information in the Web of Data". In *Linked Data on the Web Workshop (LDOW)*, 2009.
- [8] J. Zhao, A. Miles, G. Klyne, and D. Shotton. "Linked data and provenance in biological data webs". *Briefings in bioinformatics*, 2009; 10(2): 139-152.
- [9] R. Rajabi, M. Kahani, and M. A. Sicilia. "Trustworthiness of linked data using pki". In *Proceedings of the World Wide Web Conference (WWW)*, 2012.
- [10] O. Hartig, J. Zhao, and H. Mühleisen. "Automatic integration of metadata into the web of linked data". In *Proceedings of the Demo Session at the 2nd Workshop on Trust and Privacy on the Social and Semantic Web (SPOT) at ESWC*, 2010.
- [11] T. D. Nies, S. Coppens, D. V. Deursen, E. Mannens, and R. V. D. Walle. "Automatic discovery of high-level provenance using semantic similarity". In *Provenance and Annotation of Data and Processes*, 2012. Springer Berlin Heidelberg, pp. 97-110.
- [12] C. Böhm, J. Lorey, and F. Naumann. "Creating void descriptions for web-scale data". *Web Semantics: Science, Services and Agents on the World Wide Web*, 2011; 9(3): 339-345.
- [13] T. D. Nies, S. Magliacane, R. Verborgh, S. Coppens, P. Groth, E. Mannens, and R. V.D. Walle. "Git2PROV: Exposing Version Control System Content as W3C PROV". In *Posters & Demonstrations Track within the 12th International Semantic Web Conference*, 2013. CEUR-WS, pp. 125-128.
- [14] A. Stolpe, M. G. Skjæveland. "From Spreadsheets to 5-star Linked Data in the Cultural Heritage Domain: A Case Study of the Yellow List". *Norsk Informatikkonferanse*, 2011, Issue 21-23, p13.
- [15] F. Maali, R. Cyganiak, and V. Peristeras. "A publishing pipeline for linked government data". In *The Semantic Web: Research and Applications*, (pp. 778-792). Springer Berlin Heidelberg, 2012.
- [16] R. S. Xin, O. Hassanzadeh, C. Fritz, S. Sohrabi, and R. J. Miller. "Publishing bibliographic data on the Semantic Web using BibBase". *Semantic Web*, 4(1), 15-22, 2013.
- [17] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. "Describing linked datasets with the void vocabulary". W3C, March 2011.
- [18] G. Klyne, P. Groth, L. Moreau, O. Hartig, Y. Simmhan, J. Myers et. al. *PROV-AQ: provenance access and query.W3C Note*, 2012.
- [19] K. Sharma, U. Marjit, and U. Biswas. "Exposing MARC 21 Format for Bibliographic Data as Linked Data with Provenance", *Journal of Library Metadata*, pp. 212-229, 2013 published with license by Taylor & Francis ISSN: 1938-6389.
- [20] <http://www.cerl.org/resources/provenance/marc>. (Accessed on September 9, 2014).
- [21] <http://openmetadata.lib.harvard.edu/bibdata>. (Accessed on September 9, 2014).
- [22] <http://sws.ifi.uio.no/gulliste/page/dataset>. (Accessed on September 9, 2014).

Authors' Profiles

Kumar Sharma is a research scholar of the Department of Computer Science & Engineering, University of Kalyani, West Bengal, India. He obtained his bachelor degree (BCA) from University of North Bengal, India in 2006, and master degree (MCA) from University of Kalyani, India in 2009. His research interests include Semantic Web, Linked Data, and Web technologies.

Ujjal Marjit is the System-in-Charge at the C.I.R.M.(Centre for Information Resource Management), University of Kalyani. He obtained his M.C.A. degree from Jadavpur University, India. His vast areas of research interest reside in Web Service, Semantic Web, Semantic Web Service, Ontology, Knowledge Management, Linked Data etc. More than 40 papers have been published in the several reputed national and international conferences and journals.

Dr. Utpal Biswas received his B.E, M.E and PhD degrees in Computer Science and Engineering from Jadavpur University, India in 1993, 2001 and 2008 respectively. He served as a faculty member in NIT, Durgapur, India in the department of Computer Science and Engineering from 1994 to 2001. Currently, he is working as an Associate Professor in the department of Computer Science and Engineering, University of Kalyani, West Bengal, India. He is a co-author of about 90+ research articles in different journals, book chapters and conferences. His research interests include Optical Communications, Ad-hoc and Mobile Communications, Sensor Networks, Semantic Web Services, E-governance etc.

How to cite this paper: Kumar Sharma, Ujjal Marjit, Utpal Biswas, "PTSLGA: A Provenance Tracking System for Linked Data Generating Application", *International Journal of Information Technology and Computer Science(IJITCS)*, vol.7, no.4, pp.87-93, 2015. DOI: 10.5815/ijitcs.2015.04.10