Linked Data Generation with Provenance Tracking: A Review of the State-of-the-Art

Kumar Sharma, Ujjal Marjit*, Utpal Biswas

Abstract— Provenance of data items plays a pivotal role during the reuse and integration of data from the diverse sources. Determination of trust and authenticity is essential to verify various data products available on the web. Over the past few years, data publication in the Linking Open Data (LOD) cloud has been growing rapidly. Due to the absence of meta-data or provenance, data in the web starts suffering from trust and quality. Until now, lot of standard approaches has been proposed for converting legacy data to Linked Data. However, these approaches still lack a reliable method for tracking provenance of the data items. It is imperative to know the existing approaches for better understanding about how the provenance is tracked and stored. This article reviews almost preeminent approaches on tracking provenance during generation of Linked Data from the legacy data systems and presents them concisely by analyzing with various other approaches.

Index Terms— Provenance, LOD, Linked Data.

1 Introduction

PROVENANCE of a data item is a set of metadata which point to its origin from where it was derived, as well as the information about processes involved in generating the data item. In the context of database systems, provenance is defined as "*Data provenance - sometimes called 'lineage' or 'pedigree'– is the description of the origins of a piece of data and the process by which it arrived in a database*" [1]. W3C PROV-Overview[1] defines provenance as - "*Provenance is information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness*." Since long decade, it has been applied in the domain of art world to refer certain information of an art, such as, the ownership and the origin as well as the information about locations from where it was adapted. It helps in determining trust, making judgements and identifying ownership while viewing data on the web. It also helps in identifying the instructions about how to reuse data that are available on the web.

In an open environment like web, provenance provides useful guideline about a particular data product. There are consumer applications as well as publishers who wish to reuse and integrate data that are available on the web. But, without proper guideline, it becomes quite difficult in using data from third-party sources. Hence, using provenance consumers are able to know right to information, licensing information, origin and ownership of the data. Provenance has also been considered as a means to provide versioning information of data, code or any project associated with file systems.

Provenance is used in many areas. In business domains, it is used to reveal the information about manufacturing processes, compositions and usages of data products. It provides useful information about a process, enabling better understanding through input, output and used parameters. Such type of information is useful especially for engineers and scientists in large organizations and research institutions.

Provenance has recently been used in privacy and data protection scenarios [2]. As for example, personal information which is exposed into the web, securing and protecting data is a matter of concern. In such cases, provenance tends to provide infomation regarding how data is dealt with, by revealing right to information.

There are two types of provenance: *data provenance* and *workflow provenance*. *Data provenance* represents the information which is related to its history. Such as, the origin from where the data came into existence as well as other metadata. *Workflow provenance* represents the information related to processes, activities and actors that are involved in creating or processing the data item. Basically, it is a list of data flow and the activities involved in generating data item. By workflow provenance, one can find out how and when the data was prcessed and what other data were used in producing it**.**

Henceforth, provenance information is an excellent means of sharing and distributing metadata on the web, by which end-users are able to use the metadata before consuming the actual data. It provides direct perception of trust, authenticity, ownership, and privacy information on the web. Varied numbers of linked data applications have been using provenance information to improve the visibility of the data on the web. Most of the time, data published using Linked Data approach are open in nature, because of which anyone can publish any kind of data. There is no guarantee of data validity, correctness and the trustworthiness. When large organizations or government bodies publish their data, the size would be in billion of triples. This leads to problem in data management and finding trust values of the data. Hence, it is important for

- *Kumar Sharma, Dept. of Computer Sc. & Engg. University of Kalyani, Kalyani, West Bengal, India. kumar.asom@gmail.com*
- **Corresponding Author- Ujjal Marjit, C. I. R. M. University of Kalyani, Kalyani, West Bengal, India. marjitujjal@gmail.com.*
- *Utpal Biswas, Dept. of Computer Sc. & Engg., University of Kalyani, Kalyani, West Bengal, India, utpal01in@yahoo.com.*

[1] http://www.w3.org/TR/prov-overview/

publishers to track, store and disseminate provenance information of their data on the web. This led to the development of various tools and approaches to track data provenance in the Semantic Web domain.

This article reviews the work done by remarkable researchers related to legacy data to Linked Data conversion along with provenance. The structure of the paper is as follows. In Section 2 we briefly describe detail concept about provenance tracking, section 3 outlines the classification of the approaches, and section 4 presents different approaches and their comparative study. In section 5 we discuss about the reviewed approaches and section 6 concludes the work.

2 Provenance Tracking

This section tends to explain what actually the provenance tracking is and how it is tracked in a particular domain. Tracking provenance of a data item is a process of recording sequence of steps which includes the data itself, its origin, input data, output data as well the processes involved in generating, transforming, refining and recording data. Historically, notebooks were used to record these sequences of steps. It was mostly used in scientific laboratories to record the experimental results and new findings. Gradually, the demand of provenance information has been increased, which led to increase in provenance size. Finally, this way of recording provenance could not last long due to burdensome and error-prone in the result. This resulted in arising tools and softwares for recording provenance information. Though, the manual process of recording is replaced by many automated tools. But the steps and information to be recorded is still the same. These steps mainly consist of the subject matter, input parameters, the location, the computation processes, system environments, and the data items that were produced. Even the automated tools that have been developed needs human interaction at some level of tracking. For example, they may need some sort of input parameters, information about the original source, and the information about the location of the output.

Generally, provenance tracking architecture is mainly comprised of three components: *provenance collection*, *provenance storage*, and *provenance dissemination* [2] as shown in Fig. 1. Provenance collection component does the actual job of tracking. Provenance collection collects metadata from the currently running system where the data item is being produced or modified. After gathering metadata, collected information is sent to the provenance storage for representing and storing. Provenance storage follows some sort of provenance models to represent and store the provenance information. Based on provenance model, provenance collection gathers information, which is then filtered by provenance storage component. Provenance dissemination is a separate component which mainly deals about how to present provenance information to its consumer. In an open environment like web, the dissemination process should be reliable, easy to read and understandable for humans as well as it suits machine's requirement to perform query evaluation.

In the scientific domain, many software tools have been used to track data and workflow provenance. These tools use certain approaches such as literate programming, workflow management systems and environment or system capture [3]. Literate programming is the "*interactive notebook*" approach, introduced by Donald Knuth [4], which takes a description of program logic and produces executable code. Some of the literate programming tools are Dexy[2], IPython[3], and TeXmacs[4]. Workflow management systems are software systems which are used to execute and monitor scientific workflows. Workflow management systems also track provenance for the execution of various workflows. Some of the widely used workflow management systems are Taverna[5], Kepler[6], and VisTrails[7]. Environment or system capture collects the information about the environment or system (e.g. Operating systems, hardware) in which the data item was generated.

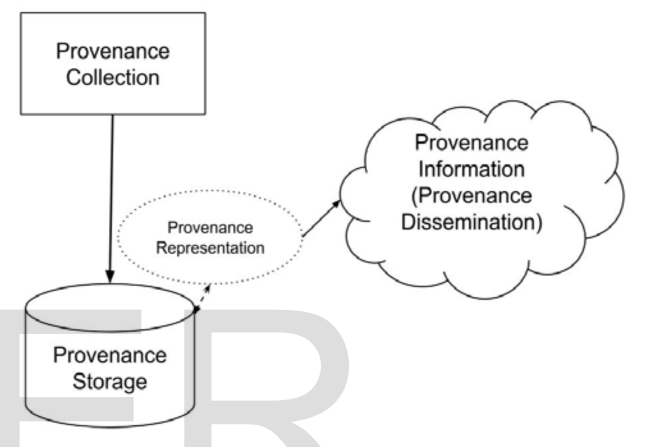
Provenance Collection
Provenance Information (Provenance Dissemination)
Provenance Representation
Provenance Storage

Fig. 1. Provenance Tracking Architecture.

3 Classification Of Aproaches

This section gives brief description of the classification terms to categorize and compare different approaches. This classification is important for understanding how provenance is tracked and stored. The classification of approaches is depicted in Fig. 2 and is described in the following:

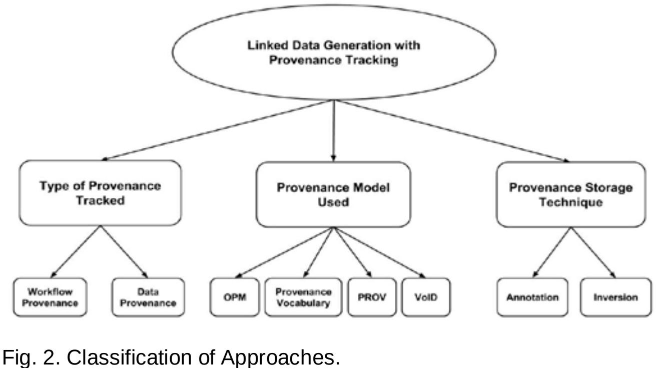

Fig. 2. Classification of Approaches.

[2] http://www.dexy.it
[3] http://ipython.org/
[4] http://www.texmacs.org/tmweb/home/welcome.en.html
[5] http://www.taverna.org.uk/
[6] https://kepler-project.org/
[7] http://www.vistrails.org/index.php/Main_Page

3.1 Type of Provenance Tracked

Provenance tracking approaches enlighten about what and how the metadata associated with a data item are captured and stored. Basically, provenance tracking is a process of recording origin & ownership, time & location, and the information about processes involved in producing the data item. Sometimes the data items are copied from one location (or database) into a new location. This step needs involvement of processes, input data, process criteria, and the actors. Here, the actor is either a human or some other process which initiates the job of data movement. There are many domains which deal with provenance, but the methodologies of tracking differs in them. Here, we categorize the approaches based on two types of provenance tracked - *data provenance* and *workflow provenance*. Data provenance is also known as *fine-grain* or *data-flow* which refers how data has moved from one location to another and its detail history [5]. On the other hand workflow provenance is known as *Coarse-grain* provenance that refers to how derived data has been calculated from raw observations [5].

3.2 Provenance Model

Provenance model provides terms and terminologies to describe and represent provenance metadata. While representing metadata one must use the terms from well known vocabularies. To deliver rich set of provenance, we need to use well known provenance models which can define data provenance as well as the workflow provenance. Provenance Models that we have considered are Open Provenance Model (OPM)[8], Provenance Vocabulary[9], W3C Provenance Model (PROV)[10], and VoID[11]. Each of these models is described briefly in the following sections.

3.2.1 Open Provenance Model

The Open Provenance Model (OPM) is a complete provenance model for describing provenance digitally of anything on the web [6]. OPM allows provenance information to be exchanged between different systems, as well as provides support for building provenance systems based on shared provenance model [6]. OPM is based on three entities – *Artifact*, *Process* and *Agent*. Goal of OPM is to represent how things came into existence and how they derived considering their different states at different time. The state of a data item is represented using *artifacts* caused by executing a process. The process is controlled by an agent. Provenance information is captured in a form of graph consisting of *entity*, *process* and *agent*. The graph depicts various dependencies and relationship among entities.

3.2.2 Provenance Vocabulary

Provenance Vocabulary [7] is a model for describing provenance of web data. This model mainly focuses on providing information about the creation of data, access-service point, creation guidelines, artifacts and actors involved in creation of the data item. Provenance Vocabulary uses a provenance graph to describe the web data which consists of three types of provenance elements viz *Actors*, *Execution*, and *Artifacts*. It states that *actors* execute some *data-creation* process to produce *artifacts*. Actors specially refer to data publishers used by some data providing services. Artifacts refer to data items, such as RDF dataset or resource. Each such provenance elements describe provenance information and holds relationship with each other.

3.2.3 W3C Provenance Model (PROV)

PROV provides provenance information about the entities, activities, and actors involved in producing a data item for assessing trust, quality and reliability of the data item [8]. PROV is mainly designed to publish the provenance information on the web [9]. PROV provides basis for representing provenance information using classes - Entity, Activity, Agent, Role, Time, Usage and Generation. These concepts take part in describing how an entity has been generated, the actor who involved in generating the entity & in performing the activity and information about used data items.

3.2.4 Vocabulary of Interlinked Datasets (VoID)

VoID is mainly used in Linked Data applications to express an RDF dataset. Using VoID, publishers can describe various metadata of the RDF dataset, such as *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). Basically, VoID helps in data discovery and cataloguing as well as helps consumers finding the right data on the web.

3.3 Provenance Storage Technique

Storing provenance is a big challenge [10]. Provenance information is represented and stored using two approaches: *Annotation* and *Inversion*. Using *annotation*, provenance is precomputed and stored separately. Whereas using *inversion,* provenance is computed on-the-fly using query based provenance services. Some criteria have been noted while storing provenance - it must be indexed to support queries [11]. It means that once provenance is stored, it must be accessible via queries, e.g. using SPARQL queries in Semantic Web technology. Also, the provenance information must persist even after the actual data is removed [11]. The big challenge is in storing provenance efficiently, where care should be taken if in case the provenance information keeps growing in size. In this category, we consider what approach has been used in storing provenance.

4 APPROACHES ON PROVENANCE TRACKING

An increasing number of data publishers have contributed their resources into the Web of Data. Provenance has also been used along with these resources, leading to evolvement of different approaches for tracking, representing and storing metadata. Many approaches provide provenance tracking system and the way of disseminating in a variety of manner. In the LOD cloud, data from any domain can be published. Such data belongs to organizations, individuals and governments.

[8] http://openprovenance.org/
[9] http://trdf.sourceforge.net/provenance/ns.html
[10] http://www.w3.org/TR/prov-primer/
[11] http://semanticweb.org/wiki/VoID

Bibliographic, government, and academic domains are one of the most widely published data. Many conversion approaches have been evolved to convert legacy data into RDF by adding more semantics and annotations on data. Such as found in [12], [13], and [14] for converting library data, sensor data and government data respectively. However, the research is still going on and the best practices are yet to emerge. Here, we present some of the preeminent approaches specially in terms of publishing legacy data as Linked Data along with tracking data & workflow provenance.

4.1 Audun et al.

A conversion from spreadsheet data to RDF is presented in [15]. Spreadsheet data contains building and other associated information such as cooking pits, burial mounds, and remnants of streets. It relies on XLWrap[12] tool for converting data. RDF data are made available on the web in the form of SPARQL end-point as well as web services which accepts SPARQL queries and returns the results in desired format. It uses two ways to track the provenance or meta-data of the dataset: using VoID description of the dataset where it is published in a URL[13]. Another way is by describing the used terminologies in the form of published vocabularies. For example, here the *?Hvor*[14] vocabulary was used to represent the address information of the buildings in the Yellow List.

The published vocabulary provides the meta-content of the properties used in a dataset. Such vocabularies follows Linked Data approach to serve data to the users, i.e., in the content-negotiated form of descriptions of the data. Such vocabularies can be developed on Neologism[15]. Neologism is a web of data publishing platform based on Linked Data principles. It helps in developing vocabularies by allowing creating RDF classes and properties, integrating them with other terminologies, and generating graph visualizations of the vocabulary.

4.2 Behshid et al.

FUM-LD (Ferdowsi University of Mashad – Linked Dataset) project is proposed by Behshid et al. [16] a framework, which converts academic data such as faculties, departments, publishers, papers (published by professors) and courses. Data of FUM is stored in the RDBMS and it is published into the web using traditional approach by choosing the appropriate entities from the FUM database and these entities are converted into RDF files, which helped in linking the FUM datasets to other LOD datasets. FUM-LD framework first generates the RDF representation of the relational entities using RDFizers[16] tool, converts the RDF entities into HTML representation using RDF2HTML, and generates the provenance of the dataset in VoID file using VoID generator. The framework uses VoID vocabulary to describe the published dataset. After generating RDF file of the dataset, the framework processes it, and generates the VoID specification of the whole dataset. The VoID file contains information about the dataset such as its subject, contributors as well as the statistical information such as number of resources, RDF triples and different subsets and link-sets of the dataset.

4.3 Boris Villazon et al.

Boris et al. [17] has proposed an approach based on "*Iterative Incremental Life Cycle Model*" to transform the legacy data into Linked Data. This approach is called single and unified method for publishing Linked Data from different domains. This approach adopts iterative life cycle model and involves following activities: *specification*, *modeling*, *RDF generation*, *Link generation*, *publication* and *exploitation*. It focuses on different domain of legacy data such as Geospatial Data, Meteorological Data, news & blogs, and Bibliographic data. There are multiple formats of legacy data such as Spreadsheet, RDBMS, MARC21 and MySQL which rely on different tools and technologies to transform legacy data to RDF. These tools are ODEMaoster[17] to convert RDBMS to RDF, NOR2O[18] to convert spreadsheet to RDF, and geometry2rdf[19] for converting geospatial data to RDF. Virtuoso Universal Server[20] and Pubby[21] are used for Linked Data publication. VoID is used to publish the metadata of the dataset.

4.4 Fadi et al.

"*Publishing Pipeline for Linked Government Data*" adopts Linked Data approach for converting raw data (government data) into Linked Data, Fadi et al. [18]. The raw data are available on different formats such as JSON, Excel, CSV and TSV. The technique is based on the Linked Government Data (LGD) Publishing Pipeline. LGD publishing pipeline is directly connected to the government catalogues using Google Refine. It uses "*CKAN Extension for Google Refine*[22]" for capturing workflow operations applied to the data. Google Refine logs all the operations applied to the data and saves in JSON files in the project history. JSON files are later extracted by OPMV (an extension to the Open Provenance Model Vocabulary) and linked to RDF data. Provenance information is represented according to the OPMV. "CKAN Extension for Google Refine[23]" is a project which allows data publishers to publish and share data on CKAN.net (an Open Data Hub). It also tracks data and workflow provenance of the dataset. Finally, the provenance information along with RDF data is shared on the open data hub, CKAN.net (http://datahub.io/).

4.5 Reynold et al.

Reynold et al. [19] presented BibBase project, which transforms bibliographic data present in BiBTeX files into Linked Data. BibBase generates the HTML as well as structured or Linked data from the BiBTeX files, performs duplicate record detection, link generation and generates the provenance information of RDF data. RDF data are stored into triple repository which is available to be queried using SPARQL. Data in

[12] http://xlwrap.sourceforge.net/
[13] http://sws.ifi.uio.no/gulliste/page/dataset
[14] http://vocab.lenka.no/hvor
[15] http://neologism.deri.ie/
[16] http://www.inf.unideb.hu/~jeszy/rdfizers/

[17] http://www.oeg-upm.net/index.php/en/technologies/9-r2o-odempaster
[18] http://www.oeg-upm.net/index.php/en/technologies/57-nor2o
[19] http://www.oeg-upm.net/index.php/en/technologies/151-geometry2rdf
[20] http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/
[21] http://wifo5-03.informatik.uni-mannheim.de/pubby/
[22] http://lab.linkeddata.deri.ie/2011/grefine-ckan/
[23] http://lab.linkeddata.deri.ie/2011/grefine-ckan/

the BibBase comes from the various sources, such as BiBTeX files. Provenance is recorded by capturing the source of each entity and links that are encoded with the entity. Provenance information helps users to know the facts about entities and fix problems such as broken links, typos, and wrong duplicate detection [19].

4.6 Steiner et al.

A mashup like browser extension tool, based on Natural Language Processing (NLP API) is presented by Steiner et al. [20]. This approach adds semantics as well as annotations to Facebook and Twitter microposts. The NLP API is based on several third party NLP APIs. This approach extracts and analyses the data from the Facebook and Twitter home pages by performing named entity extraction and disambiguation via NLP on each of the microposts. Data provenance has been tracked for understanding and optimising the result formation process. Data provenance is applied for the triples contained in named graphs by using Provenance Vocabulary[24], the HTTP Vocabulary in RDF[25], and a vocabulary for Representing Content in RDF[26]. Provenance information is recorded by tracking the source of the data, data creation process, accessed services, used data and date-time information. Provenance information has been tracked, for the following reasons [20]: to correct errors at the root of the APIs, to correct concrete errors in an RDF annotation and to judge the trustworthiness and quality of the datasets.

4.7 Vanessa et al. (QuerioCity)

QuerioCity [21] enables management of live data coming from different sources, such as, social media and physical sensors (generally urban data). QuerioCity applies modification and filtering process to legacy data, largely stored in CSV file formats, and converts into Linked Open Data along with data privacy and provenance. Furthermore, the data have been protected by detecting thwart privacy threats [21]. Linked Data approach has been used in order to provide convenient and uniform way of representation of the contents. Data privacy is applied to the datasets to protect or hide sensitive information and is applied at the dataset as well as the graph-level. Provenance is also maintained at the dataset as well as the graph level by storing *derivedFrom* relations.

4.8 Kelli et al. (LinkedDataBR).

Kelli et al. [22] presents LinkedDataBR, a project that focuses on developing a platform to facilitate the transformation and publication of legacy data (such as XML, CSV, XLS, and RDBMS) as Linked Data. The platform mainly focuses on supporting governmental data. The step involves data preprocessing (data cleansing and extraction), triplification (RDF triple generation) and linking with third party data sources. Finally, this platform provides facility to publish semantically rich dataset into LOD cloud along with provenance of the dataset. VoID vocabulary is used to describe the dataset.

4.9 Harshal et al.

Management, conversion, and tracking provenance of sensor data have been described by Harshal et al. [23]. The main objective of this project is to implement a Sensor Provenance Management System (Sensor PMS). The sensor data such as temperature, visibility, precipitation, pressure, wind speed, humidity etc. comes from various sources in legacy data format (HTML and CSV). These data have been parsed to extract sensor data and consequently they are converted to RDF using O&M2RDF converter. RDF data is stored into Virtuoso RDF store thereby producing Linked Sensor Data and Linked Observation Data. Sensor PMS mainly consists of three parts: *Provenance Capture*, *Provenance Representation* and *Provenance Storage*. Provenance capture component captures sensor provenance information, specially the spatial parameters and time stamp of the observation within the data workflow system. It uses Sensor Provenance (SP) ontology and Virtuoso RDF store for representing (modeling) and storing provenance information respectively. SP ontology is an extension of the *Provenir Ontology*[27] which helps in modeling domain-specific ontology.

4.10 Carsten et al.

Conversions of various bibliographic data into Linked Data have been presented by Carsten et al. [24]. The bibliographic metadata are present in BibTex format taken from various conference series. The metadata contains the information regarding authors & their papers, publication venues and locations. Initially, the metadata is loaded on BibTex files and then the conversion process is applied to convert into RDF. Named Graph is used to track the provenance information of RDF data, which provides information regarding from where the individual triples came from.

4.11 Jun Zhao

Jun Zhao [25] used Linked Data approach to publish Chinese Medicine (CM) knowledge as Linked Data, and further integrated this with various life science linked datasets like Entrez Gene dataset, DrugBank, DailyMed, SIDER, Diseasome, STITCH, PharmGKB and DBPedia. The CM knowledge base contains the association between medicines, genes, and diseases. All these associations are also converted into RDF. By using Linked Data approach a bridge between Chinese Medicine (CM) and Western Medicines (WM) has been made and also discussed the benefits of having open and programatic access to the data. To increase the visibility of the dataset, this approach used VoID vocabulary to publish the dataset description. Provenance of each RDF entity was described using Provenance Vocabulary. Provenance includes data creation and process related information. Further, the provenance information was published using Pubby, a linked data publication tool having support for provenance component.

5 Discussion

The survey report is shown in the Table 1. Here, we have encountered essentially two challenges - a *standard provenance*

[24] http://lov.okfn.org/dataset/lov/details/vocabulary_prv.html

[25] http://www.w3.org/TR/HTTP-in-RDF10/

[26] http://www.w3.org/TR/Content-in-RDF10/

[27] http://wiki.knoesis.org/index.php/Provenir_Ontology

tracking approach, and use of a *common representation & storage technique*. A standard provenance tracking system promotes the job of many publishers during publishing the data. Such approach tells about how the provenance information should be captured efficiently. We admit that tracking data & workflow provenance are fairly importance. Data & workflow provenance should be tracked together, since data or workflow alone can not describe the entire quality of the data. Above approaches hardly focuses on tracking data and workflow provenance together. Entity level provenance has been recorded in Reynold et al. [19] and Harshal et al. [23] by tracking the location or source of each entity. Such approach clearly tells about the source of the origin, but they sometimes lack the workflow provenance. Many consumer applications need workflow provenance, such as, the data creation process, used data items, actors and other activities. Such an approach is presented by Fadi et al. [18], in which, CKAN Extension for Google Refine has been used to track the provenance. Steiner et al. [20], Vanessa et al. [21] and Carsten et al. [24] have described provenance at the triple or Graph level. These approaches have used workflow related metadata, entity to entity relationships and the Named Graph method respectively. Audun et al. [15] used the concept of published vocabulary to describe the used properties. Such approach hardly focuses on the source or origin of the data, which only mentioned about what the property, is about.

Provenance representation using a common model is another important factor. In most of the approaches, VoID vocabulary has been used to describe RDF dataset. Nevertheless, dataset description alone can not fulfill user's requirement all the time. Resources at the triple level should also be described. Provenance Vocabulary often used in the above approaches to describe data access and creation information. However, occasionally domain specific provenance models are required to describe workflow provenance. For example, in [23] Sensor Provenance (SP) ontology (an extension of the *Provenir Ontology)* have been used. Hence, we conclude that there is no unifying model for describing workflow provenance.

Again, storage is the main concern, as in many scenarios, provenance needs excess metadata to describe the whole dataset and RDF resources. These days, a huge amount of data is being generated through different sources and applications. Consider the case, in many business applications, where the product information keeps changing by users. User changes product's behavior, stock, and different level of price information. So, the volume of the data increases on daily basis. Hence, provenance should be stored efficiently complying the criterias such as indexing, querying, and persistence even after the original data is moved or modified.

TABLE 1
DIFFERENT APPROACHES ON GENERATING LINKED DATA FROM LEGACY DATA WITH PROVENANCE

Approach	Type of Provenance Tracked	Provenance Model	Provenance Storage
Audun et al. [15]	Dataset and Property Description	VoID, ?Hvor	VoID, RDF (Annotation)
FUM-LD [16]	Dataset Description	VoID	VoID, RDF (Annotation)
Boris Villazon et al. [17]	Dataset Description	VoID	VoID, RDF (Annotation)
Fadi et al. [18]	Data & Workflow Provenance using CKAN Extension for Google Refine	Open Provenance Model Vocabulary (OPMV)	JSON, RDF (Annotation)
Reynold et al. [19]	Data provenance (Source & Link of each entity)	NA	NA
Steiner et al. [20]	Data and Workflow provenance	Provenance Vocabulary, HTTP Vocabulary	RDF (Annotation)
Vanessa et al. (QuerioCity) [21]	Data provenance using derivedFrom relation	NA	NA
Kelli et al. (LinkedData BR) [22]	Dataset Description	VoID	VoID, RDF (Annotation)
Harshal et al. [23]	Data provenance (spatial, temporal & domain parameters)	Sensor Provenance (SP) Ontology	RDF (Virtuoso RDF Store) (Annotation)
Carsten et al. [24]	Data provenance (source of origin)	Named Graph	RDF (Annotation)
Jun Jhao [25]	Dataset Description, Entity Description (Creation and Process Information)	VoID, Provenance Vocabulary	VoID (Annotation)

6 CONCLUSION

In this article, a state-of-the-art survey on flexible provenance tracking and management system during Linked Data generation has been presented. In addition to that, we have also identified how different approaches are using provenance models to represent and store the provenance. VoID has been widely applied to describe the provenance of the datasets. There are few approaches which focus on revealing provenance of data items for better visualizing and providing trust values of the data items. For doing so, there is a need of a standard provenance tracking system which can track data as well as the workflow provenance. Since, tracking provenance is still in infancy in the field of Linked Data, a standard approach for tracking data & workflow provenance is yet to emerge. Apart from this, we have summarized the different legacy data domains, use of provenance model/vocabularies, provenance representation and storage techniques.

REFERENCES

[1] P. Buneman, S. Khanna, and T. Wang-Chiew, "Why and where: A characterization of data provenance", In *Database Theory – ICDT 2001* (pp. 316-330). Springer Berlin Heidelberg

[2] C. Bier, "How Usage Control and Provenance Tracking Get Together-A Data Protection Perspective", In Security and Privacy Workshops (SPW), 2013 IEEE (pp. 13-17), IEEE

[3] http://rrcns.readthedocs.org/en/latest/provenance_tracking.html, accessed on 20 December 2014)

[4] D. E. Knuth, "Literate programming", The Computer Journal, 27(2), 97-111, 1984

[5] S. Islam, "Provenance, Lineage, and Workflows" (Doctoral dissertation, Master Thesis. Computer Science), 2010

[6] L. Moreau, B. Clifford, J. Freire et al., "The open provenance model core specification (v1. 1)", Future Generation Computer Systems 27.6 (2011): 743-756

[7] J. O. Hartig, and J. Zhao, "Publishing and consuming provenance metadata on the web of linked data", In Provenance and annotation of data and processes (pp. 78-90), 2010, Springer Berlin Heidelberg.

[8] L. Moreau, and P. Missier, "Prov-dm: The prov data model", 2013

[9] P. Missier, K. Belhajjame, and J. Cheney, "The W3C PROV family of specifications for modelling provenance metadata", In Proceedings of the 16th International Conference on Extending Database Technology (pp. 773-776), Mar. 2013, ACM.

[10] U. Marjit, K. Sharma, and U. Biswas. "Provenance representation and storage techniques in linked data: A state-of-the-art survey", International Journal of Computer Applications, 38(9), 23-28, 2012.

[11] K. K. Muniswamy-Reddy, "Deciding how to store provenance", Technical Report TR-03-06, Harvard University, 2006.

[12] Hannemann, and J. Kett, "Linked data for libraries", In *Proc of the world library and information congress of the Int'l Federation of Library Associations and Institutions (IFLA), Aug 2010*)

[13] Patni, C. Henson, and A. Sheth, "Linked sensor data", In*Collaborative Technologies and Systems (CTS), 2010 International Symposium on* (pp. 362-370), May 2010, IEEE)

[14] D.S. J. Sheridan, and J. Tennison, "Linking UK Government Data", April 2010, In*LDOW*)

[15] A. Stolpe, and 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

[16] B. Behkamal, M. Kahani, and S. Paydar, "Publishing Persian linked data; challenges and lessons learned", Journal of Mathematics, 2010

[17] B. Villazón-Terrazas, D. Vila-Suero, D. Garijo, et al. "Publishing Linked Data-There is no One-Size-Fits-All Formula", 2012.

[18] 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

[19] R. S. Xin, O. Hassanzadeh, C. Fritz, et al. "Publishing bibliographic data on the Semantic Web using BibBase", Semantic Web, 4(1), 15-22, 2013

[20] T. Steiner, R. Verborgh, J. G. Vallés, and R. Van de Walle, "Adding meaning to social network microposts via multiple named entity disambiguation APIs and tracking their data provenance", International Journal of Computer Information Systems and Industrial Management, 5, 69-78, 2013

[21] V. Lopez, S. Kotoulas, M. L. Sbodio et al. Queriocity: A linked data platform for urban information management. In The Semantic Web–ISWC 2012 (pp. 148-163). Springer Berlin Heidelberg, 2012

[22] K. de Faria Cordeiro, F. F. de Faria, et al. "An approach for managing and semantically enriching the publication of Linked Open Governmental Data"

[23] H. Patni, S. Sahoo, C. Henson, and A. Sheth, "Provenance aware linked sensor data", In Proceedings of the Second Workshop on Trust and Privacy on the Social and Semantic Web, May 2010

[24] C. Keßler, K. Janowicz, and T. Kauppinen, "spatial@ linkedscience–Exploring the Research Field of GIScience with Linked Data", In Geographic Information Science (pp. 102-115). Springer Berlin Heidelberg, 2012

[25] J. Zhao, "Publishing Chinese medicine knowledge as Linked Data on the Web", Chinese medicine, 5(1), 1-12, 2010