Provenance Representation and Storage Techniques in Linked Data: A State-of-the-Art Survey

Ujjal Marjit
C.I.R.M
University of Kalyani
Kalyani (West Bengal), India

Kumar Sharma
Dept. of Computer Sc. & Engg
University of Kalyani
Kalyani (West Bengal), India

Utpal Biswas
Dept. of Computer Sc. & Engg.
University of Kalyani
Kalyani (West Bengal), India

ABSTRACT

The Linked Open Data project has changed the world by allowing publishers to publish data of any kind as linked data and share, reuse them among other data. In Open Data each data is easily accessible and processable by machine hence one can navigate to an endless web of data sources. In the present day, many times Linked Data still suffers from trust, quality and privacy. It is requisite to endow with provenance access mechanisms that express the diverse characteristics of a dataset. In a huge volume of data-sources it is thorny to find out the trusted data and determine what a data is meant for. The Data Provenance in the Web of Data is a new technique which enabled the publishers and consumers of Linked Data to assess the quality and trustworthiness of the data. Several techniques have been emerged to represent and describe the provenance metadata in relation to the linked datasets. In this paper we appraise different techniques in this field mostly in terms of the representation, storage, and generation of provenance information of Linked Data. In addition to that we have illustrated, evaluated and identified the contemporary research challenges in this field.

General Terms
Semantic Web

Keywords
Linked Data, Data Provenance, VoID, Web of Data, URI.

1. INTRODUCTION

The traditional web only provides the web page in the front of the users as per their demand devoid of any machine processable appearance. The Web of Data is a place where all sorts of data is available and they are interlinked, shared, and reused by the data all around the world. Linked Data refers to a set of best practices to publish the structured data into the web and to interlink them with all other data so that it provides useful information by dereferencing the HTTP URIs. In Linked Data, data are interlinked with each other through a link, called RDF link, contained in the RDF document. After the Semantic Web is became a way to the Web 2.0, and as soon as the concept of Linked Data arrived and based on the four Linked Data principles described by Tim Berners-Lee [17], many organizations and business as well as academic domains has started to publish their data as Linked Data into the Web of Data, such as data from Social Media Site is made available on the Web of Data [1]. Several Linked Data applications have also been emerged that consumes Linked Data such as DBpedia Mobile [2]. Several tools have also been emerged with the intention of publishing Linked Data on the web and make them easily accessible by other data. The Web of Data is adapting its popularity day-by-day and it would be amazing if most of the information of our day-to-

day life would have linked with each other globally and directly accessible by the machine. We can see that Linked Data in the web is growing very fast as many peoples and organizations are posting data about anything into the Web of Data. The data or information in the Web of Data is accessible by the Linked Data browsers which work as same as traditional web browsers other than a way by following RDF links and querying RDF statements. Using RDF links one can navigate through different data sources such as starting from name of a person to his/her home town address, to the information about his/her friends using FOAF (Friend of a Friend) profile and to a potentially endless Web of Data sources connected by RDF links.

However, in today's dynamic world everyone needs high quality data and it is true that peoples do not have much time to navigate all the data sources and find which data is pertinent to them. Due to the openness nature of the Web, data from all around the world is growing in each second and the duplicate links are hosted at different locations on the Web. Also, it is difficult to go through a huge amount of data sources and find usefulness of them one by one. Managing provenance information of Linked Data is a new research area in the context of Semantic Web. Since the importance of provenance is very vast, the research has being made on other domains since many years. Today, Linked Data still suffers from trust and quality. Trust and quality is possible to overcome by revealing the provenance information of the Linked Data. The provenance information such as what is the originating source of the data, the access permissions, who created the data and when, will give data consumers a clear idea about the information they are looking for, so that they will be benefitted from whether or not they will meet their requirement. It would be very helpful if data publishers publish the provenance of Linked Data into the web together with Linked Dataset. However the method of publishing provenance metadata, their representation, storage location, provenance metadata generation tools are not familiar to everyone, so understanding these techniques is our main goal.

In this article we survey the existing data provenance techniques in the context of Linked Data in terms of provenance metadata representation, storage, and provenance information generation. By this it will help us to understand existing approaches, how to use them, and also the current research challenges. The structure of this paper is as follows. Section 2 gives a basic concept of data provenance, in Section 3 the basic concept of provenance representation and storage, has been discussed. A review on the current research works is illustrated in Section 4 and Section 5 concludes the paper.

2. DATA PROVENANCE

The information is a valuable asset for every people, organization, and for every business domains. As every decision are made based on the data or information perceived from various sources. To take certain decisions by analyzing data, the data should be accurate and should provide appropriate information. The provenance means lineage or tracking information about the origin of an object and the process by which it is recorded in a system. The provenance information helps us to determine trust on a particular data item and to identify the quality of the data. Yogesh L. Simmhan et al. [3] define provenance as lineage, pedigree, parentage, genealogy, and filiations. Buneman et al. [4] define data provenance in the context of database systems as the derivation of a piece of data that is in the result of the transformation step or description of the origin of a piece of data and the process by which it arrives in a database. Data provenance also helps to find out the errors by tracking its history behind its creation such as data creation process, creation time, who was involved in the creation process and what materials were used during creation. For example, before buying any product customers would like to know the provenance information such as where the product has came from, who made it in what conditions and using what materials, what are the materials that effects this product and under what conditions should this product be kept to avoid from damage so that the customers can trust upon it and make use of it or can determine whether or not the data is valuable for them. Consider an object whose information or knowledge is stored in the database and in case, for example, of a failure or damage of the object, people might be interested in to track down its creator and the ingredients the object is made up of, in such cases the query is fired on the database and retrieved from it. This concept is mainly called traceability or tracking data back. Provenance is also associated with its process(es) or workflow(s) and for which there are two types of provenance- data and workflow provenance. Workflow provenance refers to the record of the entire history of the derivation of the final output of the workflow, such as recording of the software programs, hardware, and the instruments used in the experiment, [4], and data provenance is something like recording the events related to data. The basic reason to use provenance is that it helps in determining trust on agents and web resources, because it is possible to get fake, suspicious, illegal copyright while using data from third-party in today's loosely connected network, so data provenance can help to solve these issues.

3. PROVENANCE REPRESENTATION AND STORAGE

Representation and Storage of provenance information in Linked Data are two main challenges. Representation of provenance metadata describes how to represent the provenance metadata terms and it should be easy and understandable which can collect enough information. After representing provenance metadata we need to store provenance information in an efficient way. Tope Omitola et al. [5, 6] also argue that Provenance Representation and Provenance Storage are the two major challenges in the provision of provenance information of Linked Data

3.1 Provenance Representation

To the best of our knowledge, provenance is represented by two different approaches viz *annotation method* and *inversion method,* also discussed in [3, 6]. The two approaches are described below:

Annotation Method: In annotation method the provenance information is pre-computed. This method requires a separate metadata to store the collected provenance information hence this method keeps the provenance explicitly as metadata and are ready to use.

Inversion Method: In inversion method the source data is obtained based on the properties that some derivation process can be inverted (working backward). For example queries and user defined functions in databases that is inverted automatically or by explicit functions. The derivation queries are used to create an inverse query that operates on the output data and auxiliary data to identify the source data [3]. Here the provenance is computed when needed.

From the above two approaches the annotation method gives more flexibility to represent varieties of provenance metadata and the provenance need not be computed on-the-fly as in inversion method. There is one limitation in annotation approach that while describing the provenance information in detail sometime it may exceeds the size of original data. One way to support annotation method is to use VoID which is discussed in section 3.2.

3.2 Provenance Storage

Provenance information can be stored within the same data storage system where the data it describes is stored, or it can also be stored in different storage location with other metadata [3], such as VoID. VoID (Vocabulary of Interlinked Datasets) is an ontology used to describe the elements of the RDF datasets. Generally, VoID acts as a bridge between the publishers and the users of the RDF datasets since it advises the publishers how to describe and discover the metadata of their datasets. It also tells the users of RDF datasets about what a dataset is about and how to access it. VoID is mainly used to express General metadata, Access metadata, Structural metadata, and links between datasets. It is based on two main concepts as described below:

Dataset: The dataset is a collection of RDF triples which is published and maintained by a single provider. void:Dataset is used to model a dataset. e.g. :xDataset a void:Dataset.

Linkset: VoID also allows the description of RDF links between the two datasets. An RDF link is an RDF triple whose subject and object are described in two different datasets. A linkset is a collection of such RDF links. void:Linkset is used to model a linkset. e.g. :xDataset_geolocations a void:Linkset .

Basically VoID describes various metadata of the dataset and the provenance can be inserted or stored using this metadata description. Keith et al. [7] describes everything about VoID and has presented some use cases how VoID can be applied to describe RDF datasets.

4. CURRENT APPROACHES

Here we summarize various techniques and approaches to represent, store, and generate the provenance information of Linked Data. Recently, Edoardo Pignotti et al. [17] has discussed that provenance plays an important role in revealing the information about data, since whatever we publish in Web of Data is actually a data not an information. So the provenance is a context which supports the assessment of quality, reliability, and trustworthiness of information [18]. A set of provenance principles have been proposed which enable the publishers to bring out the provenance information of the Linked Data to make data more transparent as well as provide trustworthiness and eminence of data. These principles act as

guideline framework for us to publish provenance of Open Data. The provenance principles are:

- *One Star Provenance*: Publish the provenance of data on the web whatever format (e.g. plain text).

- *Two Star Provenance*: Publish provenance as structured data (e.g. database, spreadsheet, XML)

- *Three Star Provenance*: Use URIs to identify individual elements within the provenance record.

- *Four Star Provenance*: Link provenance record to other provenance records using RDF.

Provenance information is published in several formats however for the Web of Data it will be supportive if we publish Three Star and Four Star provenances as both of them facilitate applications and users to rely on information within the provenance record. Adhering to these principles we summarize following research works.

Tope Omitola et al. [5, 6] has discussed on the subject of the use of provenance information into Web of Data and defined a vocabulary called VoIDP. VoIDP is a lightweight provenance vocabulary, an extension to the VoID vocabulary, which enables publishers to portray the provenance information of their linked dataset. Solely VoID ontology is not provided adequate information to the queries such as "when a data item was created and who were involved in the creation process?". VoIDP provides classes and properties, which is designed by reusing existing vocabularies such as Provenance Vocabulary [10], The Time Ontology in OWL [14], and the Semantic Web Publishing Vocabulary (SWP) [15]. The classes and properties is used by means of data publishers to describe the provenance information of their dataset enabling services to answer urbane user queries so that the publishers can improve the quality or the value of their data The list of classes and properties of VoIDP vocabulary is depicted in the following tables as Table1 and Table 2. Classes of VoIDP vocabulary are actor, provenance and data item. Different types of properties with their descriptions are elaborately explained in the Table 2.

Table 1: Classes of VoIDP vocabulary (Adapted from Tope Omitola et al. [5, 6])

| Class	Description
Actor	An Actor performs an action on a particular data item. Reused class from Provenance Vocabulary.
Provenance	A container class for the list of DataItem(s).
DataItem	This class models the item of data that are put under provenance control.

Table 2: Properties of VoIDP vocabulary (Adapted from Tope Omitola et al. [5, 6])

Property	Description
activity	Specifies that a particular dataset has some items under provenance control,
item	Specifies the item under provenance control
originatingSource	Item's original source
originatingSourceURI	The URI of the item's original source
originatingSourceLabel	The label text used to de- scribe the item's original source
certification	Contains the signature element
swp:signature	Represents the signature of the dataset
swp:signatureMethod	Specifies the signature method
swp:authority	Defines the authority of the relation- ship between the item under provenance control and the dataset publisher.
swp:valid-from and swp:valid-until	The valid start and end dates of that (authority) relationship,
processType	Specifies the type of transformation or conversion procedure.
prv:createdBy	Specifies the actor that executes an action on the item that is being recorded
prv:performedAt	Date of transformation
prv:performedBy	The URI of the actor that performs the recording of the provenance activity on the item

Olaf Hartig in [10] has elaborately discussed regarding the provenance of web data and proposes a suitable provenance model that captures both information about web-based data and information about the creation of data which is used by publishers to depict their web data. Using this model author aims to assess the data qualities such as accuracy, reliability and appropriateness as well as aptness. The Provenance information of a Web data is described using a provenance graph whose nodes are three types of provenance elements viz Actors, Execution, and Artifacts. Each element describes unambiguous provenance information and holds relationship with each other. For example, an Artifact (Artifact represents the end products such as the dataset) is a result of an Execution performed by an Actor. In case of a data creation process the elements involved are Data Creation (Execution), Data Creator (Actor), and Data Item (Artifact). This scenario gives us the information about who created the data item using what data creation process. The Actor or Data Creator (Human or Non-Human) rivet in the Data Creation (Execution) to yield the Data Item (Artifact). Each of these provenance elements can have attributes such as Data Creating Service, Data Creating Device, Creation Time, Creation Guideline etc. which in turn represents precise provenance information. In a similar way, the provenance

information is possible to describe for the Data Access dimension. Furthermore, the information about digital signature is also feasible to depict as the provenance graph whose elements are Integrity Assurance, Digital Signatures, Public Keys, and Signers. In this case the Integrity Assurance is a verification process which verifies the Digital Signatures of the Signer.

Li Ding et al. [12] presented an approach on evaluating data quality by tracking the provenance of RDF graph. The information such as "where an RDF graph came from", "who has created an RDF graph" would benefit the users to trust on RDF data. Currently this approach provides the provenance information at the document level such as the document URL of the RDF documents, creator and inter-document
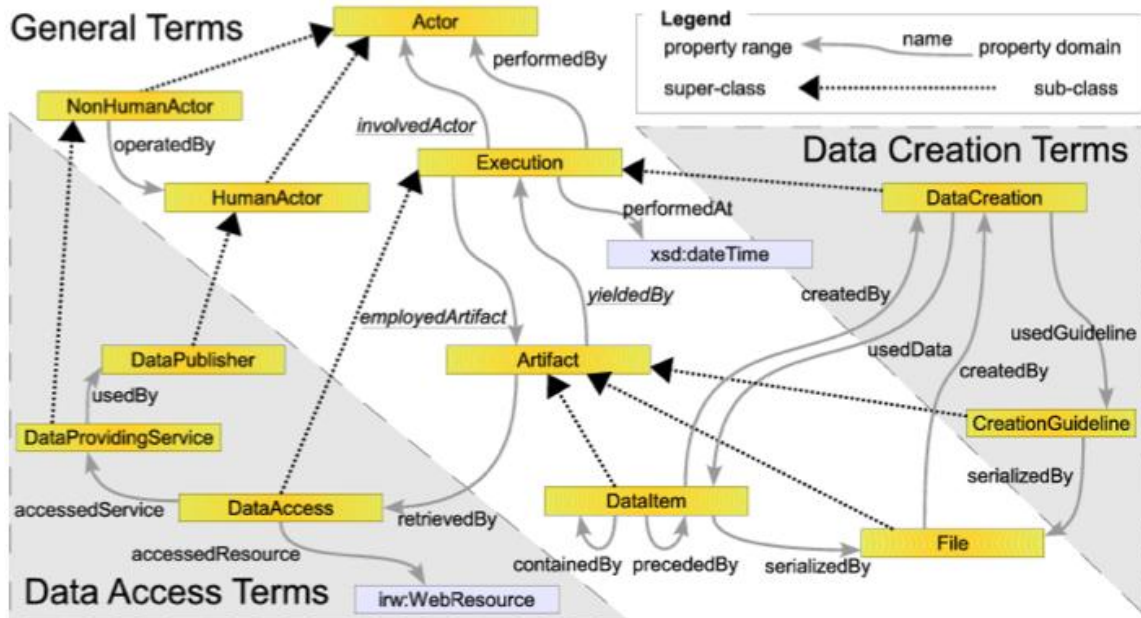
Figure 1: Classes and Properties of Provenance Vocabulary (Adapted from Haritg [11])

Olaf Hartig and Jun Zhao in [11] outlined approaches on providing quality aware Linked Data into the web, how to consume such data to know the provenance of a specific data item which will help to identify the obsolete, poor quality, and wrong information. A recommendation has been made to generate the provenance of Linked Data in the equivalent way as we publish the Linked Data into the Web of Data. Author proposes a vocabulary called Provenance Vocabulary which allows the publishers to describe provenance of data on the web. After having the metadata terms from Provenance Vocabulary the provenance information of the dataset are stored using VoID which is based on two dimensions viz Data Creation and Data Access, as identified in [10]. Provenance Vocabulary consists of three parts: General terms, terms for Data Creation, and the terms for Data Access which are depicted in the Figure 1. The classes of general terms are Actor, Execution Artifact, DataItem (subclass of Artifact), File (subclass of Artifact) etc. and the properties include yieldedBy, performedBy, performedAt, operatedBy, etc. Using general terms we can describe provenance information such as an Artifact is yieldedBy an Execution and it is performedBy an Actor which is performedAt a particular time. The classes of data creation dimension are DataCreation, CreationGuideline and the properties are createdBy, usedGuideline, usedData etc. Using Data Creation terms we can again describe provenance information as, a DataItem is createdBy a DataCreation using usedData (source data). Similarly, we can describe provenance information for Data Access dimension whose classes includes DataPublisher, DataProvidingService, DataAccess and properties are usedBy, retrievedBy, accessedService, and accessedResource.

relationships. Authors propose to track the provenance information of an RDF graph by RDF Molecules, (a new level of granularity which lies between RDF document and triple level granularity as shown in Figure 2), which are the finest and lossless components of an RDF graph, into which the graph is decomposed without loss of information [12]. Authors also argue that tracking the whole RDF document is too coarse grained since it could contain irrelevant information and also the RDF triples are too fine level granularity of RDF graphs with blank nodes [12].
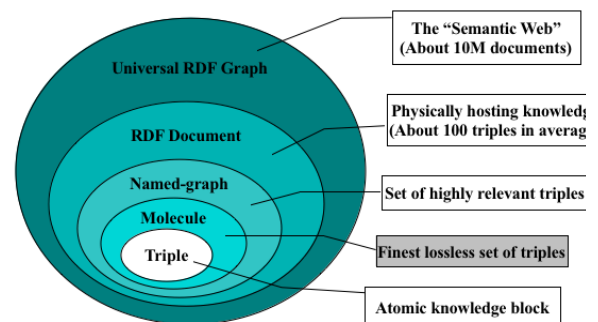
Figure 2: Introducing the Molecule granularity of the Semantic Web (Adapted from Li Ding [12])

Andreas Harth et al. [13] put forward a social model of provenance for tracking provenance information of people or groups of people. The provenance information is associated with the originator (such as person or organization) of a data

items. Since most of the information on the web is created by people or groups of people, hence obtaining provenance information, particularly in the context of social networking, would help users to ask over sophisticated queries regarding people. Here the provenance information regarding peoples is recorded by tracing the URIs of the data sources via the Domain Name System (DNS).

Jun Zhao et al. [16] used the concept named graph (a named graph is an RDF graph which is assigned a name in the form of URI [16]) to deal with the provenance information concerning the links between the data items from various sources. In a biological data-web data items from diverse sources are linked with each other, however to get the detail information about the links from databases users had to go through a information gathering process navigating between different databases. To resolve such issues authors applied the named graphs to explain the provenance information of the links. Using named graphs authors were able to record the provenance information such as the evidence of links, when a link was created, when it was updated, who has created a link, whether two links are identical or not, whether a link is obsolete or not, etc. They have used the terms from existing vocabulary such as dc:creation and dc:creator from the Dublin Core Metadata Element Set to define metadata terms. In addition to that authors have added a *dw (http://www.datawebs.net/)* namespace to spell out certain properties of their own.

Olaf Hartig [19] argues with reference to a lack of suitable method to endow with reliability and trustworthiness of RDF data on the Web. He has suggested an approach to trust or construct judgment on the RDF data that have been published on the Web and propose a trust model that associates each RDF statement with a trust value, a number that lies between [-1, 1]. Depending on the trust values one can trust an RDF statement where -1 represents an absolute belief, 1 represents absolute disbelief, and values other than -1 and 1 represents a lack of belief/disbelief or unknown trust. Here the knowledge about trust is depend on information consumer and is determined by Trust Assessment Methods. Author has focused on Opinion-based (opinion of information consumers of the data) and Provenance based methods. Provenance-based methods rely on provenance information of the data. The provenance information that affects trust decision, so far identified by author are publisher of the dataset, creation method and creation time of the dataset, and publisher and publication time of possible original sources [19]. To make use of trusted RDF statements which are associated with a trust value author has introduced some Access Methods such as an extension to the RDF based query language SPARQL called tSPARQL a trust-aware query language, (available at http://trdf.sourceforge.net/tsparql.shtml), which allows access to the trustworthiness of RDF data.

4.1 Provenance Representation

From our research and understanding it shows that the representation and storage of the provenance information are the two main research challenges that are facing the present semantic web environment. We argue that storing the provenance information in the data file itself or the same data storage system may create some issues like the management of the provenance information, mainly the scalability issue, would be difficult if the provenance metadata grow in size and what if the size of the provenance metadata is larger than the data it describes? However, if we store the provenance information with other metadata as in VoIDP [5, 6] then once the provenance metadata is defined and stored then the change

or update should be reflected in provenance information if something is updated or inserted into the original data. The data may always be updated or introduced new sort of information, hence the versioning of the provenance information should be kept in mind. However, VoIDP provides a rich set of classes and properties, as we mentioned in Table 1 and Table 2. We can use them to describe various provenance events of data items inside VoID document.

Provenance Vocabulary [11] which is based on Data Creation and Data Access dimensions allows us to describe the web data provenance and relationship between different types of actors, executions and artifacts, however Tope Omitola et al. [5, 6] argue that encoding at the triple level may cause the provenance information to be very large in case of a large dataset.

RDF Molecule [12] used to track the provenance information of RDF graph at appropriate level of granularity. This approach separates out the irrelevant information contained in RDF document and blank nodes contained in RDF triples. Using RDF Molecule [12] we have to access the objects of RDF triples which may take considerable amount of time to access the object in case when the dataset contains a huge number of triples.

Provenance is also described using *named graph* as discussed in [16] which may overcome most of the issues however this concept is not much popular and also it may require variation in current RDF specification. To understand detail about how *named graph* works we need further research.

Social Model [13] of Provenance captures provenance information that is associated with originator of the data items such as who has created a data item and who posted it. Such kind of provenance is useful in Social Networking context but it may not capture the complete information about data items.

Trust Model [19] caters a novel approach of discovery trustworthiness of data on the web. Since Linked Data suffers from trustworthiness of RDF data so we anticipate the said approach will overcome the problem by making web of trusted data. However, the trust- aware system requires strong trust assessment methods to determine the trust values and the efficient management of trust ratings.

Provenance information generation tool is also an important part of provenance metadata management and we realize that there are lacks of such tools to do the same. Users might seek provenance information services which can visualize the provenance information that can meet their requirement. In [8] authors have mentioned about the extension of several Linked Data publishing tools such as Triplify, Pubby, and D2R Server. These extensions would help publishers to publish provenance related metadata along with their original data. It is expected that using these tools we will be benefited from publishing the web of data as well as their provenance information generation. In near future we are going to study and talk about them in our research work.

5. CONCLUSION

In this paper we have presented a state-of-the-art survey of existing approaches and techniques of managing Provenance of Linked Data mainly in terms of their representation, storage, and generation of provenance information. We have also illustrated how the provenance information assists to identify the quality and trusted data and discussed different approaches to represent the provenance information. It is depicted a set of provenance principles proposed by Edoardo Pignotti et al. [18] for publishing provenance information of

Open Data which are similar to the Linked Data principles described by Tim Berners-Lee [17]. The provenance principles act as guideline framework for us to publish provenance of Linked Data and most of the current research works is done in the same way what these principles suggest. We need further research work how adhering to these principles and using best approaches of representation and storage we can publish the provenance information of Linked Data and make data more transparent. We have gone through some of the research works and understood their approaches and evaluated how they are useful in different contexts. Nowadays, representations, storage of provenance information are two main open issues in the field of provenance. As far as our knowledge is concerned, there are lack provenance information generation tools for efficient discovery of provenance information to the end users. However, authors [8] have discussed about extension of Linked Data publishing tools such as Triplify, Pubby, and D2R Server. In our future work we will find out how these tools are helpful in terms of publishing and generation of provenance information. Keeping these open issues as key points we will also experiment on this domain that fulfills some of the current research issues.

6. REFERENCES

[1] Uldis Bojars, Alexandre Passant, Richard Cyganiak, John Breslin, "Weaving SIOC into the Web of Linked Data". Presented at WWW 2008 Workshop Linked Data on the Web (LDOW 2008), Beijing, China, April 2008.

[2] Christian Becker, Christian Bizer, "DBpedia Mobile: A Location-Enabled Linked Data Browser". In 1st Workshop about Linked Data on the Web (LDOW2008), April 2008.

[3] Yogesh L. Simmhan, Beth Plale, Dennis Gannon, "A Survey of Data Provenance Techniques", Indiana University, Technical Report IUB-CS-TR618, 2005.

[4] Peter Buneman, Sanjeev Khanna, Wang-Chiew Tan, "Why and Where: A Characterization of Data Provenance". In Proceedings of the 8th International Conference on Database Theory (ICDT). Springer, Jan. 2001.

[5] Tope Omitola, Landong Zuo, Christopher Gutteridge, Ian C. Millard, Hugh Glaser, Nicholas Gibbins, Nigel Shadbolt, "Tracing the Provenance of Linked Data using voiD". *The International Conference on Web Intelligence, Mining and Semantics (WIMS'11).*

[6] Tope Omitola, Nicholas Gibbins, and Nigel Shadbolt. "Provenance in Linked Data Integration". Future Internet Assembly, 16-17 December 2010, Ghent, Belgium.

[7] Keith Alexander, Richard Cyganiak, Michael Hausenblas, Jun Zhao. "Describing Linked Datasets, On the Design and Usage of voiD, the Vocabulary Of Interlinked Datasets". LDOW2009, 2009.

[8] Olaf Hartig, Jun Zhao, and Hannes Muhleisen. "Automatic Integration of Metadata into the Web of Linked Data". Proceedings of the 4th International Workshop on Semantic Web Enabled Software Engineering (SWESE) at the 7th International

[9] Uldis Bojars, Alexandre Passant, Richard Cyganiak, John Breslin, "Weaving SIOC into the Web of Linked Data". Presented at WWW 2008 Workshop Linked Data on the Web (LDOW 2008), Beijing, China, April 2008.

[10] Olaf Hartig ,"Provenance Information in the Web of Data". Linked Data on the Web Workshop at WWW (2009).

[11] Olaf Hartig, Jun Zhao, "Publishing and Consuming Provenance Metadata on the Web of Linked Data". In Proceedings of the 3rd International Provenance and Annotation Workshop (IPAW), Troy, New York, USA, Jun. 2010.

[12] Li Ding, Tim Finin, Yun Peng, Paulo Pinheiro da Silva, and Deborah L. McGuinness. "Tracking RDF Graph Provenance using RDF Molecules". Technical Report TR-CS-05-06, UMBC (2005).

[13] Andreas Harth, Axel Polleres, Stefan Decker. Towards A Social Provenance Model for the Web. Proceedings of the Workshop on Principles of Provenance. (2007).

[14] J. R. Hobbs and F. Pan. An ontology of time for the semantic web. ACM Transactions on Asian Language Processing (TALIP): Special issue on Temporal Information Processing, 3(1), 2004.

[15] Christian Bizer. "Semantic Web Publishing Vocabulary (SWP) User Manual". Available at http://www4.wiwiss.fu-berlin.de/bizer/wiqa/swp/swp-usermanual.pdf

[16] Jun Zhao, Graham Klyne and David Shotton, "Provenance and Linked Data in Biological Data Web". LDOW2008, Beijing, China, April 22, 2008.

[17] Christian Bizer, Tom Heath, Tim Berners-Lee. "Linked Data – The Story So Far", International Journal on Semantic Web and Information Systems (IJWIS), Vol. 5, No. 3, pp. 1-22, 2009.

[18] Edoardo Pignotti, David Corsar, Peter Edwards. "Provenance Principles for Open Data". In Proceedings of DE2011, November, 2011.

[19] Olaf Hartig, "Trustworthiness of Data on the Web". In Conference Proceedings of the STI Berlin & CSW PhD Workshop, Berlin, Germany, September. 2008.