State of the art Attribution with regards to Open Data

Abdul Saboor

Research Assistant at Freie University Berlin,
Department of Computer Science,
Takustr. 9, 14195 Berlin, Germany.
abdul.saboor@fu-berlin.de

Abstract

In this paper, our aim is to highlight the importance of attribution and emphasis on state of the art attribution process, key elements of attribution and current issues in attributions, tracking the source of data, which are based on the discussion about provenance of data on the web. It also provides guidelines for organizations and individuals about the mechanism of publishing more transparent data on the web. Provenance plays an important role in terms of open data assessment and it enables to evaluate the data attributes such as quality and trustworthiness. On the basis of provenance information, the quality and completeness of open data can be assessed through data analysis.

Keywords

Attribution, State of the Art, Provenance, Open Data, Metadata, Trust, Authentication, DC Vocabulary, OPM.

1 Introduction

Attribution is a process that enables readers to get information about the source of particular information or data. It is analyzed that who is the originator of this particular information such as a department or authors job title or his position in an organization and also the purpose of the information. Attribution mainly emphasis on a particular platform, such as a website in which an information or datum is given in various formats and the publisher of this information needs to be established in order to track the source of information for authentication and validity.

When the information or data originated from a proper source then it will be more authenticated and justifiable for any readers, viewers and listeners and they can use that information for analytical or research purposes, they can also integrate some more information and can published the updated version. Similarly when open data is collected from the web then it should be analyzed: where does it come from? Who is the owner of this particular type of open data? It should be ensured that the data belongs to an authenticated organization so that data can be used for further analysis or research purposes. On the web, mostly open data comes from various sources all around the world and in some cases it is difficult to ensure that the attribution for that datum when the source is not provided or is ambiguous in terms of data validity.

2 Attribution

In context of data, attribution is identifying the source where that data come from in order to ensure the data authenticity it is analyzed that who produced that particular kind of data, e.g. person name, title, organization name, etc. The main objective of attribution is to define the appropriate characterization of documents that captures the source of generated information and creation method of authors and or organization.

2.1 Definitions of Attribution

"Attribution often involves identifying the author or source of the written material or a work of art " [1].

"Attribution is the acknowledgement of the use of someone elses information, data, or other work " [2].

3 Data Attribution

According to Kotarski R, Reilly S, Schrimpf S, Smit E, Walshe K (2012), the aim of Data Attribution is to acknowledge the data creators and indicating availability of data. In attribution of data, there are some considerations due to particular properties of datasets that are as follows:

- **Granularity:** It mainly emphasis that what features inside the datasets are being referred to.
- **Versioning:** In the context of dynamic or regular updated data, which version needs to be attributed.

The actions from those who have an active role in data reuse have the potentials to improve data attribution:

- ◎ **Location of data:** Provide a persistent link such as Digital Object Identifiers (DOI) and accession numbers to ensure sustainable access to the attributed data.

- ◎ **Acknowledging creators:** It has to be ensured that the credit is given to those who deserve it.

4 Importance of Attribution (state of the Art)

There are various features specified by Christine L. Borgman 2012, in which we can analyze the importance of attribution [3].

4.1 Social Practice

It is important in social perspectives that data must be described in such a way so that it can be discovered in appropriate form. For those people who are devoting their efforts in making data discoverable they should get the credit of creating, analyzing, sharing and making data available and useful for other users. In order to get credit, it has to be associated the names of individuals and organizations with specific units of data. Attribution is mainly associated with the notion of contribution or the person who contributed than with author. However, attribution implies the social responsibility to give credit to that person who deserve for that credit. When we write articles, reports and books we give the references of other publications and also give the evidences on which they are based in order to attribute our sources.

4.2 Usability

Attribution needs to be consider in terms of data usability, as the data which we have for our analysis or study that is in different forms such as samples, artifacts, lab paper notes, these should also be referenced. There is a wide range of capabilities amongst the actions of people or machines which they wish to perform on digital objects that are; open, read, evaluate, interpret, compute upon, combine, re-use, annotate, these should be accommodated through a process of attributing data.

4.3 Discoverability

It is the ability to determine the existence of a set of data objects with specified attributes or characteristics. The attributes of interest that includes; the creation of data, date of creation, method of creation, a description about the objects contents and its representation. Discoverability also involves some relevant aspects of data such as levels of quality, certification and validation by third parties. Discoverability depends on data description and data representation, and those tools and services, which are used to search for data objects. The most common approach which is used now a day to discover data and information regarding to it; it is web search engines such as Google, and data descriptions are reachable through standard web protocols.

The introduction of Semantic web technologies and associated search engines, the locations of interested datasets based on semantic web contents have become to access.

Datasets are discoverable as long as someone keeps them available somewhere on web. Some data can be discoverable only for short term and other data will be kept at least until the associated reports are published. Data attribution practices and standards may vary depending on the period of time that data are expected to remain available for viewers.

4.4 Relationship

Multiple types of data often have relationship to each other regarding to an organization or an institution, by providing context, calibration and comparisons. Data attribution mechanism facilitates linking of related datasets and make able to refer the other group of data items as well as the individual items.

4.5 Policy

Data attribution has the policy components, and many stakeholders are concerned with scholarly information policy, including funding agencies, investigators, publishers, data repositories, universities and students. Everyone has a policy concern, they must ask about what policy, whose policy and what kinds of policy. Many funding agencies or organization have established such kind of policies, the specifics are varies depending of organizations policies such as the Economic and Social Research Council in UK, National Institute of Health and National Science Foundation in USA, European Union, European Commission. These requirements can change to become more precise about who is to receive what kind of attribution for what kinds of data contributions.

4.6 Provenance of Open Data

Provenance is important to reference the current version of data. In computing, provenance is the ability to track all transformations from original state. Provenance describes the history and life cycle of data where this data originated, in other words the source of information where that information initiated such as the entities and process which involved in producing or delivering that artifact. The provenance of information is vital for establishing whether that information is trustable or not. It is also crucial to determine how to integrate the diverse information sources and how to specify the details about originator or how to give him credit when re-using that information. The Web is an open environment of information where users find information from various sources that are questionable and sometimes the information is contradictory. People make trust judgements based on provenance that may or may not be explicitly offered to them. Researchers in the Semantic Web field will need an explicit representation of provenance in order to make trust judgements regarding to that information which they use. For instance, when developing new Semantic Web

Applications with a massive amount of data via linked open data community, the information about the origin of data becomes an important factor.

Furthermore, provenance is mainly concerned with a broad range of sources and their uses. The concept of provenance in terms of its origin and preceding ownership is important to determine the data authenticity. In the scientific context, data depends on collection and the methods that are use for pre-processing and also the validity, which is determined through experimental results that are based on how each analysis step was performed [4].

4.6.1 Definitions of Provenance

"Provenance of a resource is a record that describes entities and processes involved in producing and delivering or otherwise influencing that resource. Provenance provides a critical foundation for assessing authenticity, enabling trust and allowing reproducibility. Provenance assertions are a form of contextual metadata and can themselves becomes important records with their own provenance" [4].

"Provenance refers to the source of information, including entities and processes, involving in producing or delivering an artefact"[4].

"Provenance would include information about the creation and publication of web resources as well as information about access of those resources, and activities related to their discussion, linking, and reuse" [4].

4.6.2 Provenance and Metadata

The term Metadata is structured information that describes, locates, or otherwise makes it easier to retrieve, use, or manage an information source. Metadata can also be defined as Structured data about an object that supports functions associated with the designated object. The structure implies a systematic data ordering according to a metadata schema specifications; the object can be any entity or form, which contextual data can be record; and the associated functions can be activities and behavior of the object. Provenance information and Metadata are important for the scientific analysis where software needs to be able to identify the datasets are appropriate for particular analysis and then interpret the new. The descriptive metadata becomes the part of provenance only in that case when someone specifies the relationship metadata in order to drive an object. For instance, a file can have a metadata property that states the file size and that cannot be considered provenance information unless it does not relate to how that file was created. When the same file has metadata regarding to its creation date then it can be considered as provenance relevant metadata. Provenance is often represented as metadata but it is not necessary that all metadata is provenance [6].

4.6.3 Provenance and Trust

Trust is a most common term which is widely used and has many definitions. When we talk about establishing trust while analyzing an object or an entity which contains its origin and authenticity. Trust is also associated with provenance but it is derived from provenance information and it is a typical subjective judgement, which depends on the use and context of information. The provenance is the mechanism for trusts procedures and approaches on the web. But in terms of authentication of data, provenance guides you to establish the trust on data while providing the data source information. Presently the available mechanism for authentication of a source is the verification of an identity or the access to a resource such as digital signatures. Provenance information can also be used for authentication purposes; for instance, the creator of a document can provide an evidence (e.g, IDs, or Digital Signature, or Source code, etc) that can be verified through various legal authorities or by third parties [6].

5 Why should the attribution be used?

According to Kotarski R, Reilly S, Schrimpf S, Smit E, Walshe K (2012), the research process involves building on, reusing and critically evaluating the published data of evidence or information related to a particular field. Attribution allows author to acknowledge the original work and give an opportunity or facility to the readers into the background of the process that can guide to their current work. There are a number of reasons are given why authors attribute their particular work. Hanney et al (2005) give a valuable summary of various motivations, for instance, refuting or supporting a work simply noticing it, reviewing a work because it contains information that is being applied in some way. These motivations can be equally applied to all works, regardless of their formats. So the reason of attributing data is to acknowledge the original creator of data and keep the track record datasets while following the attribution pattern.

5.1 Main Reasons of Data Attribution

According to Thomas Lee 2002, there are various reasons on behalf of these reasons attribution is performed and these reasons for attribution are as follows;

5.1.1 Data Quality

The purpose of attribution is to ensure that we are having the proper sources of datasets. These sources that are provided by some organisations and their individuals can be verified through queries in order to ensure the accuracy or correctness and completeness of dataset for a specific fact. When the data source is reputable then we are much more likely to accept the accuracy of the facts. For answering a particular query, knowing the sources used in each step of the query process helps to verify the accuracy of the answer set accordingly. The correctness of the query answer also

depends upon whether the sources used to evaluate query conditions, such as the figures of unemployment ratio in 2012 of a particular city, are accurate and up to date. There are multiple ways of deriving an answer, multiple sources for specific values, in order to ensure whether there are contradictory results, are all the ways reinforcing confidence in a specific answer.

5.1.2 Previous work for verification and reuse

The intellectual work, insights in a research publication often include the data interpretation and allowing others to access the underlying data, it can allow for alternative interpretations and hypotheses to be derived. Providing access to data also allows the researchers to check mistakes and find inconsistencies. For this purpose researchers need to know where to find these datasets and how to get the access, and there is sufficient information about the datasets need to be provided that how that was generated.

5.1.3 Maintaining the research record

As the new research is based on the understanding what has done before, the research come from various authors. Attribution of existing work is important to give an overview that how this area has advanced, the way of understand a subject has been changed over the time, the datasets are equally important aspect of this record.

6 Elements of Data Attribution

There are different elements or ways of data sets attribution that are as follows [7]:

- ⊙ **Dataset Name:** Specify a name for each dataset that represent to a particular organization (e.g., filename/document name such as EU Coral Reef dataset, etc.).
- ⊙ **Authors Name and Contact Details:** Specify the name of the author of data and contact details (e.g., organisation name and address, telephone number, e-mail address, etc.). It needs to be mentioned that who is the creator or owner of data or the primary person who is responsible for that data, and also it can be specified the secondary source the person who is currently handling the data. This information is mandatory in terms of analysing and reviewing the datasets for answering the raised questions in events or meetings.
- ⊙ **Data Description:** Provide enough description, which reflects the contents of datasets accurately.
- ⊙ **Data Formats:** Specify the list of various data formats that are supported such as csv, xml, xls, html, pdf, txt, rdf/n3, etc.

⊙ **Data Handling Rules:** Describe the particular data handling rules or policies which apply on the data and that must be needed to follow. For instance, some datasets have a license such as Creative Commons CC0 1.0 Universal (http://creativecommons.org/publicdomain/zero/1.0/). If there is any restriction then it also needs to specify the particular conditions of data handling.

⊙ **Data Access Methods:** It has to specify the access method that how someone can get access to the data either via a URL (if know then specify it) or an API (if relevant then mention the API such as web-service SOAP, web-service REST), or any database then it is also important to describe the product and version such as SQL Server 2005 or 2008, Oracle Database 11g or Release2 or Standard/Enterprise Edition, MySQL Enterprise Edition, Cluster CGE or Workbench 5.2, etc.).

⊙ **Dataset Size:** It is also necessary to specify an estimated sized of dataset (e.g., less than 10 MB or more than 100 MB, 1 GB or greater than such as 1.3 GB, etc.). If the dataset size is exactly known then specify, e.g. 135.9 MB in a particular format, or in case of database enter the number of records.

⊙ **Data Time-period:** Specifying the time period for the data, which also tells that the data is for this particular time period, for example, data between 2001-2010, data for 2011, etc.).

⊙ **Data Status:** It explains that how often the dataset is updated, either it is updated on hourly, daily, weekly, monthly or annually basis. If it is real time data, then specify the average data rate such as 100 kbps, 1 mbps.

⊙ **Data Factors:** It specifies the name of factors which involve in the dataset, for example, time, year, square meter, etc.

⊙ **Data Collection Methods and Interpretation:** Explain the level of difficult that may exist in while collecting and/or interpreting data by users. There are three levels of obtaining data that are; easy, semi-challenging, and challenging. In the context of easy, the data is structure and expressed in digital format and it is available online via a URL. In semi-challenging context, the data is semi-structured and expressed in digital format. In challenging, the data is unstructured and/or unavailable in digital form and that needs to bring in understandable form.

⊙ **Data Availability:** It is explained that the data already exist and is available for users and if it is not then how that data become available on web.

⊙ **Language of Data:** It is specified that the data, which is available is in one language or support some other languages. If it does support other languages then it has to be mentioned means provide the list of supporting language.

7 Current Issues and Challenges in Attribution

Our aim is to express the key issues relating to data attribution and there are key elements which need to be considered in order to make attribution process more appropriate and for tracking of data. Data attribution is the key to the successful adoption of data sharing by researchers and that can help to address relevant issues that need to be tackled if best practice in data attribution is implemented.

7.1 Granularity

According to Alex Ball and Monica Duke (2012), Datasets attribution is relatively a little more complicated. A dataset can be a part of several files; each file contains many tables, records and data points. There are some additional subsets that can be used, such as features and parameters. For authors, the practical solution is to list datasets at whatever level of granularity has been chosen by host repository for assigning the identifiers. If the repository provides identifiers at several levels of granularity, then the finest-grained level that fulfill the requirements of attribution should be used.

7.2 Contributor Identifiers

Alex Ball and Monica Duke (2012) explain; every contributor has some uniqueness in their organizational activities, if there is two different institutions have common names then give them a unique identifier to each contributor, to be used in connection with data contributions. There are two schemes being developed specifically for the attribution which are as follows:

1. The Open Research and Contributor Identifier (ORCID) is a scheme specifically for academic authors. This scheme has got support from over 200 organizations, including academic publishers. The intention is to maintain a registry of IDs, and each associated with a researcher profile and a list of publications that a researcher has contributed. The registry will allow the profile to be linked with identifier and profile from other schemes such as Scholar Universe.

2. The International Standard Name Identifier (ISNI) scheme is a standard for registering Public Identifies such as people, personas, legal entities involved in the creation or distribution of intellectual property. It is a broader scheme allowing organizations to be identified and their individuals as well. ISNIs take the form of 16-digit number; each identifier is supported by a metadata record containing details such as name, date of birth, position, title of creations and a URI for further information.

7.3 Micro-Attribution

Micro-Attribution is a way of crediting contributors in a more compact way, in order to keep process manageable. It can be used to credit people or organisations whose

contributions do not fit the roles of data creator or compiler: for instance, those who carry out or implement immediate data processing steps. Instead of providing a traditional attribution to the data collection paper that is associated with each contribution, a table is created that provide the list of each contribution and the responsible agent. The standard identifiers for both contributors and contributions are used to abbreviate the entries, and a table is included in the documents supplementary data.

8 Current Implementation Issues

There are two current issues for data repositories that how to deal with both manual and automatic uses of attribution, and how to deal with dynamic datasets [8].

8.1 Manual and Automatic use of Attribution

According to Alex Ball and Monica Duke (2012), It is good practice for the URL in terms of data attribution to lead to a landing page for the dataset, rather than to initiate or direct download dataset. The landing page will enable the readers to ensure they have located the right dataset, to familiarise themselves for research context and supporting documents, to consider the licence terms prior for downloading and switching for more recent version of data if it is required. Landing pages can also help to create a better user experience between available datasets through direct access and those available through referred access.

The most of the parts of data are processed by software tool; it can help to accelerate progress and software tools are able to retrieve data through same URL. Software tools provide support to readers; they may wish to be selective to download with regard to versions and formats, to select individual files or subsets of data, and to avoid with that data which has some licence restrictions. These types of use cases require that the URL actually return the machine-readable equivalent of landing page.

In addition, human and software have different requirements for landing page for the dataset. In order to satisfying them there is a way that both would be to embed the metadata intended for software tools as a Resource Description Framework (RDF) with the human-readable web page. This action can also be performed by using, either RDFa (Resource Description Framework-in-attributes), or HTML5 micro-data. There is an alternative method used for serving them that is content negotiation. This is possible where the web server keeps several kinds of representation of a resource; when a web client requests a resource, the server sends back the representation that best matches the clients preferred content type as stated by the Accept HTTP header. In this situation, the web server keeps the landing page of dataset an HTML web page for human readers and an RDF/XML for software tools. While archives and repositories are mainly consist of the information which they provide to readers on their landing pages the description metadata, an attribution, a link to the data files or some instructions on how to get access to them, and licence terms.

8.2 Versioning

Alex Ball and Monica Duke (2012) describe, an important feature of attribution system is that a reader be able to identify and retrieve the exact same resource that the author used. This is critical in terms of data and it can significantly change the conclusions that are drawn from a dataset. There is also a possibility for more versions from which to choose, since the data from various stages of processing can be made available in different versions. The data repositories should also be ensured that different versions are attributed independently with their own identifiers.

The problem arise when repositories have to deal with rapid changes in datasets, and it is bit different problem depending on whether the dataset is frequently revised, and data are continuously improved or updated periodically. In order to keep the various versions manageable there are two potential approaches that the data repository can take: time slice and snapshots. Our aim of description is mainly concern that how datasets are presented to users as attributed sources.

9 Approaches used for the Attribution

There are some approaches which are used to support the attribution of data such as Dublin Core Vocabulary which expresses the attribution using metadata that is an organization or a user can play a particular role in the life of the resource, and in such a role which is being played that resource was affected by some method on some date. The Open Provenance Model is a fundamental graph model of provenance which expresses the relationship of processes between data objects.

9.1 Dublin Core Vocabulary

Dublin Core (DC) Vocabulary approach provides a vocabulary for expressing the resources. DC strength relies on shared usage across different repositories and organizations. A common vocabulary can be used in order to search or index across the distributed resources, and distributed application use Dublin Core terms for communication about resources. DC emerged from the library and the archiving communities but it is common in web-focused research by allowing the annotation of web-accessibility and other resources with agreed upon metadata [9]. Dublin Core consists of a set of qualifiers and a core set of metadata elements which make it possible to interpret the elements in a semantic way. In the context of attribution, a subset of elements and qualifiers can be employed, for instance, there are terms which are used for the creator of a resource, for its publisher, and also for the dates of its publication. A typical Dublin Core metadata statement consists of three fundamental points which are:

1. An identifier for the resource being described
2. A term from the Dublin Core vocabulary

3. The annotation value

For instance, This Paper has a Title of State of the art attribution with regards to Open Data. This information can be encoded in many data formats but the Resource Description Framework form is the most appropriate for this type of information (triples), and DC provides URI versions for each of its terms and can be used as RDF properties [10].

9.2 Open Provenance Model

The Open Provenance Model (OPM) is a process in which the data is being produced or transformed into a new state, and it can represent the provenance of one or more data items from an old to a new state. OPM is a basic graph model for provenance which provides the description of provenance about the graph whose edges denote the primary relationships between the occurrences represented by the nodes. This type of structure allows OPM graphs to explain how multiple events conducted to produce some sort of data (independently or serially), and shows how one part of data derived from another part. OPM classifies nodes into three parts: Artifacts, Processes and Agents. Artifacts are the parts of data of fixed value and context that possibly represent an entity in a given state. Processes are performed on artifacts in order to produce another artifact. Agents indicate the entities which are controlling the processes, such as users. Edges can also have annotations for providing the information on how one occurrence caused another [11]. The main purpose of OPM is to support the assessment of various data qualities such as reliability, accuracy and timeliness.

10 Vocabularies that support attribution

There are different kinds of vocabularies which can be used to describe the attribution information for the RDF data. These vocabularies are described below with their related classes and their properties are also defined through these vocabularies to the elements of the provenance model. A popular standard is used to represent the general purpose metadata: Dublin Core Metadata Terms which are available as RDF schema [12]. The following properties are defined by this schema and can be associated with a resource in order to describe the provenance information.

- ⋆ **dcterms:creator.** The creator property is an entity and the primary responsibility is making the resource [12]. This property can be used to acquire information about the data creators of a data item. Analyzing data about the creator can give further information which can help to derive a more precise provenance graph.
- ⋆ **dcterms:source.** This property describes the source of a resource is a related resource from which the described resource is derived [2]. By using this property it is possible to create provenance elements which are associated as source data with a data creation element.

⋆ **dcterms:modified.** This property specifies the date in which a resource has been changed. The modification of data item as a data creation which makes a new modified version of original data item. By using this property, the creation time attribute associated with this data creation can be set.

⋆ **dcterms:publisher.** The publisher of a resource is an entity responsible for making the resource available. This property can be used to acquire information about a provider of an information resource whereas the actual information provider (data provider service, service provider and data publisher) remains uncertain.

⋆ **dcterms:provenance.** This property makes a link to a resource with a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity and interpretation [12]. It is difficult to use such a provenance statement during the creation of a provenance graph.

The Friend of a Friend (FOAF) vocabulary [13] provides the properties and classes to describe entities such as persons, groups, communities, organizations, agents, Etc. FOAF provides the descriptions for obtaining the basic information about the actors such as names, e-mail address, identifying online accounts, group membership. In addition, FOAF contains the property foaf : maker and its inverse property is foaf : made and the purpose is to relate the described entities to the resources which are made by entities. These properties can be used for identifying the creator of data and the data items.

The semantically Interlinked Online Communities (SIOC) ontology [14] expresses information from online communities. This ontology associates the SIOC items such as comments, blog post, and e-mail messages to the users that are identified by their online accounts. The properties which describe the provenance-relevant information are as follows:

⋆ **sioc:has-creator , sioc:creator-of , sioc:has-modifier , sioc:modifier-of.** These properties relate a SIOC item to a user who created it or who modified it. The creator and modifier referenced to SIOC ontology based description and that are data creating entities.

⋆ **sioc:has-owner , sioc:owner-of.** These properties describes the ownership of a SIOC item. This information indicates the relation to a data publisher and data provider, and it might be used to set the corresponding attribute of the entity in a provenance graph.

⋆ **sioc:earlier-version , sioc:later-version , sioc:next-version , sioc:previous-version.** These properties relate different version of SIOC item with each other, and these properties can be used to create relationship between the object and a provenance graph.

- ⋆ The Semantic Web Publishing (SWP) vocabulary [15] provides the information about provisioning of data. It is possible to represent the attitude of a person to a RDF graph through SWP and it supports two attitudes such as: (1) Claim the graph is true, (2) Without a comment quoting the graph on its truth. These statements for the truth can be used to derive the data publisher or data creator relation to data provider or created artifacts.
- ⋆ The Web Of Trust (WOT) schema [16] provides information about documents that uses the public key cryptography tools to sign documents. It refers that signatures are individually encoded in the dedicated document, and these digital signature described with WOT sign the information source.
- ⋆ The Ontology Metadata Vocabulary (OMV) [17] includes the properties for creator, distributor, reviewer, and the creation and modification dates.
- ⋆ The Changeset Vocabulary [18] describes the changes to RDF-based resource descriptions.

11 Related Work

There is a significant amount of research done from several years in this particular area of provenance of data and there are some models are also defined. The main purpose of that research is to analyze the provenance information about the data which comes from many sources over the web and it has to discover that the particular kind of data where it comes from, who is the creator of that data and after utilizing that data by some organization or individuals then who modified it and provide that data into new state, simple drive the relationship between data publisher and data provider. In addition, Olaf Hartig 2009 presents an abstract provenance model for the Web of Data. This model description is similar to the OPM but without distinction between the artifacts and processes. The author differentiates the two dimensions of Web Data Provenance that is the provenance about the creation of the data or how it originates from a particular source, and the access to that data, in other words the method and source to retrieve that particular data or sets of data. That work further describes the potential Semantic Web encodings for provided provenance model [19].

Moreover, Simmhan et al. [20] provides the classification of the provenance characteristics. The authors are distinguishing the two approaches that are: the data-oriented approach and the process-oriented approach. The data-oriented approach expresses the data items and the process-oriented approach provides the information about the processes which are performed to generate the data. There are several questions raised by Bunemann et al. [21] for data provenance in modern era of Web. The authors highlight the three major issues in context of data provenance which are:

1. Obtaining provenance information

2. Citing components of digital library such as the component of a document in another context

3. Ensuring the integrity of citation whilst assuming that cited databases evolve

Hausenblas et al. [22] presented another aspect of Web Data Provenance. The authors differentiate the sources of Web Data based on the way these sources represent the RDF data, and these sources may contain RDF data in a non-serialize form (e.g. in-memory, in-store) or the random data can be in serialized form. Sources with serialized data can be:

1. RDF model-compliant and standalone

2. RDF model compliant and embedded

3. Non-compliant to RDF model

12 Conclusion

The purpose and process of attribution is very much crucial almost everywhere either educational environment, organisational environment public or private and individual basis, the data which is taken from a source should be attributed with description such as authors name, title and the date of publish because this information will make the provenance more authenticated and trustable. An attribution mechanism can keep the originality of work and it will also appreciate the creators to introduce more innovative work while integrating their ideas to previous work done by other authors. It might serve to enhance the reputation of organizations worldwide through this mechanism and can build their confidence and also trust in people. Attribution is a formal way of publishing information or data on the web. When certain steps are performed in a systematic way such as the information about the data who is the creator (name and title) of that data and through which institute or organization this data is published and at which date that data is released on web. Through these steps it will be much easier to do provenance because all the information about data is interlinked which drives the relationship between the creator and modifier.

References

[1] Tony Rogers, Attribution Definition, How to use attribution in a new story. *http://www.vocabulary.com/dictionary/attribution and http://journalism.about.com/od/writing/a/attribution.htm.*

[2] The Mind Wobbles, Attribution vs Citation: Do you know the difference? *http://themindwobbles.wordpress.com/2009/07/10/attribution-vs-citation-do-you-know-the-difference/* . July 2009.

[3] Christine L. Borgman, Why are the attribution and citation of scientific data important? Report from Developing Data Attribution and Citation Practices and standards. *An International Symposium and Workshop*, January 2012.

[4] W3C Website, What is provenance? *http://www.w3.org/2005/Incubator/prov/wiki/What Is Provenance*, Modified at November 2010.

[5] W3C Website, A working Definition of Provenance. *http://www.w3.org/2005/Incubator/prov/wiki/What Is Provenance AWorkingDefinition of provenance*, Modified at November 2010.

[6] W3C Website, Provenance, Metadata, and Trust. *http://www.w3.org/2005/Incubator/prov/wiki/What Is Provenance Provenance.2C Metadata.2C and Trust*, Modified at November 2010.

[7] Edzard Hofig, Jens Klessmann, Nils Barnickel (Fraunhofer), Open Innovation mechanism in Smart Cities, Revision: A, v1.6, July 2011.

[8] Alex Ball and Monica Duke (2012), How to Cite Datasets and Link to Publications, Revised June 2012.

[9] D.G. Campbell, The use of Dublin Core in web annotation programs.*In proceeding of the International Conference on Dublin Core and Metadata Applications*, Florence, Italy 2002, pp105-110.

[10] Simon Miles, Mapping Attribution Metadata to the Open Provenance Model, .Future Generation Computer Systems 27 (6), Kings College London, UK, pp. 806811, 2011.

[11] Dublin Core Metadata Initiative Usage Board, DCMI Metadata Terms. *http://dublincore.org/documents/dcmi-terms/*, January 2008.

[12] Olaf Hartig, Provenance information in the Web of Data, Humboldt University Zu Berlin. *In proceedings of the 2nd Workshop on Linked Data on the Web (LDOW2009)*, April 2009.

[13] D. Brickley and L. Miller, FOAF Vocabulary Specification. *http://xmlns.com/foaf/spec/*. November 2007.

[14] U. Bojars and J. G. Breslin. SIOC Core Ontology Specification, Revision 1.30. *http://rdfs.org/sioc/spec/*, January 2009.

[15] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler, Named Graphs, Provenance and Trust. *In Proceedings of the 14th International World Wide Web Conference*, ACM Press, pp613-622, May 2005.

[16] D. Brickley. Web of Trust RDF Ontology. *http://www.w3.org/tr/rdf-schema/*, February 2004.

[17] R. Palma, J. Hartmann, and P. Haase. OMV - Ontology Metadata Vocabulary for the Semantic Web, v2.4. *http://omv2.sourceforge.net/*, January 2008.

[18] S. Tunnicliffe and I. Davis. Changeset Vocabulary. *http://vocab.org/changeset/schema.html*, March 2006.

[19] Li Ding, James Michaelis, Jim McCusker, and Deborah L. McGuinness. Linked Provenance Data: A Semantic Web-based approach to interoperable workflow traces, Elsevier, Future Generation Computer Systems, Vol.27, October 2010.

[20] Y. Simmhan, B. Plale, and D. Gannon. A Survey of Data Provenance in e-Science. SIGMOD Record, Computer Science Department, Indiana University. Vol. 34, Issue No. 3, p3136, ACM, September 2005.

[21] P. Buneman, S. Khanna, and W. C. Tan. Data Provenance: Some Basic Issues. *In Proceedings of the 20th Conference on Foundations of Software Technology and Theoretical Computer Science (FST TCS)*, p87-93, Springer, December 2000.

[22] M. Hausenblas, W. Slany, and D. Ayers. A Performance and Scalability Metric for Virtual RDF Graphs. *In Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW) at ESWC*, June 2007.