Tracing the Provenance of Linked Data using voiD

Tope Omitola
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
tobo@ecs.soton.ac.uk

Landong Zuo
The Stationary Office
United Kingdom
Landong.zuo@tso.co.uk

Christopher Gutteridge
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
cjg@ecs.soton.ac.uk

Ian C. Millard
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
icm@ecs.soton.ac.uk

Hugh Glaser
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
hg@ecs.soton.ac.uk

Nicholas Gibbins
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
nmg@ecs.soton.ac.uk

Nigel Shadbolt
Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
nrs@ecs.soton.ac.uk

ABSTRACT

In the open world of the (Semantic) web, a world where increasingly diverse materials from disparate sources of different qualities are being made available, an automatic mechanism for the provision of provenance information of these sources is needed. This paper describes **voidp**, a provenance extension for the **void** vocabulary, that allows data publishers to specify the provenance relationships of their data. We enumerate voidp's classes and properties, and describe a use case scenario. A wider uptake of voidp by dataset publishers will allow data consuming tools to take advantage of these metadata providing consumers with the origin, i.e., the provenance, of what is being consumed.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Linked Data; D.2.8 [**Software Engineering**]: Semantic Web

General Terms

Web Science

WIMS'11, May 25-27, 2011 Sogndal, Norway

Keywords

Lineage, Provenance, Linked Data

1. INTRODUCTION

How do you identify what is good? Most social interactions require matching human needs on one hand, with quality or taste, e.g. hunting for a reliable mechanic, looking for an interesting book, sifting through potential investments, judging the merits of proposed government policies, etc. On the Web, a user may be confronted with a potentially large number of diverse data sources of variable maturity or quality, and selecting the high quality data that are pertinent for their uses may be difficult. They would like to have mechanisms to automatically determine whether a web document or resource can be used, based on the original source of the content, the licensing information associated with the resource, and any usage restrictions on that content, etc. In cases of mashed-up content, i.e. content that is the result of aggregation of information from a wide variety of sources, it would be useful to ascertain automatically whether or not to trust it by examining the processes that created, processed, and delivered it.

Provenance, also known as lineage, describes how an object came to be in its present state, and thus, it describes the evolution of the object over time. For example, in the business world processes rely on human activities that may not be predicted in advance, and information exchange is heavily based on ephemeral and unstructured data, such as e-mails or attachments, where the content is unstructured and needs discovery. Visibility of such end-to-end operations is required to manage compliance and business performance, and, hence it becomes necessary to develop techniques for

tracking and correlating the relevant aspects of business operations.

In e-Science, provenance of the process of creation of an experiment's dataset may include information on the instruments used to make measurements, the identities of individuals and agencies responsible for creation, and the standards used to define the dataset's contents. By knowing such details, it is possible in many cases to make inferences about quality. For example, knowing the identity of the instrument used to acquire measurements often allows the user to make meaningful estimates of their accuracy. Provenance also serves another useful purpose by providing feedback, for example, if serious errors are found in the data, it might be possible to link them to specific faults in the production process. This provenance data can be used in some cases as documentation permitting repeatability of experimental results, and therefore the independent confirmation of findings as regards the experiment.

There is an increasing trend in governments and industry in the adoption and usage of ontologies and semi-structured data to publish their datasets bringing about the growth in datasets published in linked data format, and a growing interest in connecting these datasets together. Linked Data is a style of publishing data on the Web that emphasises data reuse and connections between related data sources. This growth and interest can be seen in the Linked Data community which aims at making data freely available to everyone and to extend the Web with a data commons by publishing various open data sets as RDF and by setting RDF links between data items from different data sources. Applications are then built on these open data commons, an open environment that contains data from a wide variety of different sources that can be meshed together and re-used in a number of powerful ways. Applications built on this open data platform, of diverse sources, will need mechanisms to enable the automatic identification of the processes of creation and the origins of these sources. Further information that might be useful include: what source data has been used for creation; where, how, and when has the data (or the source data) been retrieved from the Web; who is responsible for accessed services, etc.

2. PROVENANCE

There are two major research strands of provenance prominent in the literature: **data** and **workflow** provenance. In the scientific enterprise, a workflow is typically used to perform complex data processing tasks. A workflow can be thought of as a set of procedure steps, computer and human, that one enacts to get from the starting state to the goal state. Workflow provenance refers to the record of the entire history of the derivation of the final output of the workflow. The details of the recording vary from one experiment to another. It may depend on the goals of the experiment, or the regulatory and compliance procedures, and a number of other things. It may involve the recording of the software programs, the hardware, and the instruments used in the experiment.

Data provenance, on the other hand, is more concerned about the derivation of a piece of data that is in the result of a transformation step. It refers to a description of the origins of a piece of data and the process by which it arrives in a database. This paper considers both workflow and data provenance.

If we have a set of workflow items or data items we want to provide provenance information for and our provenance metadata is in semi-structured form as a collection of objects, where each object is atomic or complex. The value of an atomic object is of some base type (integer, (URI) string, image, sound, etc). The value of a complex object is a set of $< attribute, object >$ pairs, where an attribute is any string drawn from a universe $\mathcal{A}$ of attribute names. Our provenance metadata can be modelled as a graph, where the nodes are objects, the edges are labelled with attributes, and leaf nodes have associated atomic values. The graph has a root, i.e. a distinguished object, with all other objects accessible from it.

Formally, our semistructured data is $G =< who \cup < V, E, r, v >>$ where who is the identifier of the actor that performed the provenance operation, V is the set of nodes partitioned into complex and atomic nodes $V = V_c \cup V_a$, the edges are $E \subseteq V_c \times \mathcal{A} \times V$, $r \in V$ is the root, and $v : V_a \rightarrow \mathcal{D}$ assigns values to atomic objects and $\mathcal{D}$ is the universe of atomic values.

So, given an evolving universe $\mathfrak{U}_t$ of G at time t, and a finite subset of G, $\{G_i\}_{1 \leq i \leq M} \subset \mathfrak{U}_T$ from the universe of G at a time T, the Provenance questions asked of a data item x, that is one of the nodes of G, are the following:

1. When was x derived (when-provenance), i.e. what is the lowest value of t for which $\mathfrak{U}_t$ contains an item y_a which has contributed to the evolution of x in $\{G_i\} \subset \mathfrak{U}_T$,

2. How was x derived (how-provenance), i.e. what were the first $y, (y = y_1 \ldots y_n)$, in the chain that culminated in the current value of x in $\{G_i\}$,

3. What data was used to derive x (what-provenance), i.e. which $y, (y = y_1 \ldots y_n)$, in $\mathfrak{U}_t$ for $t \leq T$ contributed to the evolution of x in $\{G_i\} \subset \mathfrak{U}_T$,

4. Who carried out the transformation(s) from whence x came (who-provenance), which who was attached to $v : V_x \rightarrow \mathcal{D}$.

Both (2) and (3) can both be elements of *why-provenance*[6].

The provenance problem is gaining more prominence in the Semantic Web community. The advent of Linked Open Data[1] has made it a primary concern of data consumers to consider whether the data is usable based on its provenance[2]. Reasoners in the Semantic Web need explicit representations of provenance of the information they use in order to decide what assertions and axioms to use. Provenance is also important in determining trust on agents and web resources.

3. RELATED WORK

There are many surveys of existing work on provenance from workflows [2] and database[9] research communities. At times, provenance has been conflated with trust and some work in the trust research communities [10] incorporated provenance as part of their work. There have been

[1]http://linkeddata.org

[2]"Provenance is the number one issue we face when publishing government data as linked data for data.gov.uk". *John Sheridan, UK National Archives*, **data.gov.uk**, February 2010.

some work on the quality assessment of data that have addressed the issues of provenance[5]. There is also the Open Provenance Model[15] which allows the characterisation of the dependencies between "things", and it consists of a directed graph expressing such dependencies. The nodes represent the artifacts, processes, and agents, while the edges have predefined semantic relationships that depend on the type of the nodes. It is not light-weight but can be used to describe part of the provenance relationships that is a concern of a dataset publisher.

In this work, we focus on the Semantic Web, especially Linked Open Data. Berners-Lee's "Oh yeah?" button [3] was meant to challenge the origins, i.e. provenance, of what is being asserted and request proofs, by directly or indirectly consulting the meta-information of what is being asserted.

Named graphs [7] are models that allow entire groups of graphs be given a URI and provenance information can be attached to those graphs. The Semantic Web Publishing Vocabulary (SWP)[4] is an RDF-Schema vocabulary for expressing information provision related meta-information and for assuring the origin of information with digital signatures. It can be used within the named graph framework to integrate information about provenance, assertional status, and digital signatures of graphs. An RDF graph is a set of RDF triples, therefore an RDF graph may contain a few triples or very many. The Named Graph framework does not give a good control on the granularity of the collection of data items to attach provenance to. In this work, we do use some elements of the SWP.

The Provenance Vocabulary[11] provides classes and properties enabling providers of Web data to publish provenance-related metadata about their data. The vocabulary provides classes, called Artifacts, Executions, and Actors, that can be used to specify provenance for data access and data creation, at the triple level. An Actor performs an Execution on an Artifact. In the Provenance Vocabulary, there are different types of actors that perform different types of executions over diverse types of artifacts. Although encoding at the triple level is fine-grained and lets provenance data be attached to a triple, a big dataset may contain a large number of triples, and encoding at triple level may lead to the provenance information be much more than the actual data.

In the linked data world, data are usually collected together and provided as datasets. The provision of the provenance information of datasets' elements is an interesting problem.

4. PROVENANCE OF DATASETS

A sizeable proportion of the content of the semantic web is built by hand, including ontologies, linked data, mashups, etc. This means that many assertions were crafted by people based on their understanding of the domains being modelled. To create these assertions, the developer usually consults documents and sources, makes some assumptions, and integrates information. It would be useful to record, in as much detail as possible, what were the original sources consulted, what pieces seemed contradictory or vague, which were dismissed, what additional hypotheses were formulated in order to complement the original sources, etc. However, this kind of information is not captured in current practice. Ontologies, assertions, and resources lack such records to provide rationale for their design, and as a result it makes it hard for others to reuse those ontologies and data. Depending on the underlying computer system, different techniques can be used to represent provenance.

4.1 Provenance Representation

There are two major approaches to representing provenance information, and these alternate representations have implications on their cost of recording and the richness of their usages. These two approaches are:

- The Inversion method: This uses the relationships between the input data, working backwards (hence the name "inversion"), to derive the output data, giving the records of this trace. Examples include queries and user-defined functions in databases that can be inverted automatically or by explicit functions [17]. Here, information about the queries and the output data may be sufficient to identify the source data,
- The Annotation method: Metadata of the derivation history of a data are collected as annotation, as well as descriptions about source data and processes. Here, provenance is pre-computed and readily usable as metadata.

While the inversion method is more compact than the annotation approach, the information it provides is sparse and limited to the derivation history of the data. The annotation method, however, provides more information that includes more than the derivation history of the data and may include the parameters passed to the derivation processes, the post-conditions, etc.

For this work, we chose the annotation method as it gives richer information of the data and the data set we may be interested in. We adopted the voiD (Vocabulary of Interlinked Datasets) [1] vocabulary to describe the provenance information of the data we are interested in. voiD is an RDF based schema to describe datasets. With voiD, the discovery and usage of datasets can be performed both effectively and efficiently. There are two core classes at the heart of voiD:

1. A dataset (*void:Dataset*), i.e. a collection of data, which is:
   - published and maintained by a single provider,
   - available as RDF,
   - accessible, for example, through dereferenceable HTTP URIs or a SPARQL[3] endpoint
2. The interlinking modelled by a linkset (*void:Linkset*). A linkset in voiD is a subclass of a dataset, used for describing the interlinking relationship between datasets. In each interlinking triple, the subject is a resource hosted in one dataset and the object is a resource hosted in another dataset. This modelling enables a flexible and powerful way to state the interlinking between two datasets, such as how many links there exist, the kind of links, and who made these statements.

After choosing the representation mechanism, the next questions to ask are:

- what data points would generate the data salient to our provenance needs, and

[3]http://www.w3.org/TR/rdf-sparql-query/

- what is the minimal unit of a dataset to attach provenance to.

A small dataset may contain a minimal amount of data items, while a large dataset may contain millions of data items. Although it would be more appropriate to attach provenance to whole collection of data items that make up the dimensions of a dataset, however this decision of the minimal unit to attach provenance to is left to the dataset publisher. Provenance representation has an affect on the scalability of its storage.

4.2 Provenance Storage

Provenance information can sometimes be larger than the data it describes if the data items under provenance control is fine-grained and the information provided very rich. The inversion method may prove to be more scalable than the annotations method. However, one can reduce storage needs by recording data collection that are important for the operational aspects of the dataset publisher's business.

Provenance can be tightly coupled to the data it describes and located in the same data storage system or even be embedded within the data file, as advocated in tSPARQL [12]. Such approaches can ease maintaining the integrity of provenance, but make it harder to publish and search just the provenance. It can also lead to a large amount of provenance information needing to be stored. Provenance can also be stored by itself [26] or with other metadata. We chose to store the provenance information with the other metadata of the dataset(s) (using voiD description files).

5. VOIDP : PROVENANCE EXTENSION TO VOID

In [8], a linked data(set) publisher was advised to reuse terms from well-known vocabularies wherever possible, and one should only define new terms one cannot find in existing vocabularies. Reusing existing vocabularies takes advantage of the ease of bringing together diverse domains within RDF, and it makes data more reusable. By reusing vocabularies, the data is no longer isolated nor locked within a single context designed for a single use. We adhered to this advice and have made use of the following ontologies:

- Provenance Vocabulary[13],
- The Time Ontology in OWL[14],
- The Semantic Web Publishing Vocabulary[5],

In addition, the namespace for voidp[4] is:
`@prefix voidp: <http://purl.org/void/provenance/ns/>.`
We intend **voidp** to be an easy to use ontology, and thereby designed it to be light-weight with few classes and properties. The classes are:

- `Actor`: Here, we reuse the Actor class in the Provenance vocabulary to specify an entity or an object that performs an action on a particular data item (or a data source or data set),
- `Provenance`: this class is a container class for the list of DataItem(s) we are putting under provenance control,
- `DataItem`: this class models the item of data we put under provenance control.

The properties are:

1. `activity:` this property specifies that a particular dataset has some items under provenance control,
2. `item`: specifies the item under provenance control,
3. `originatingSource`: the item's original source,
4. `originatingSourceURI`: the URI of the item's original source,
5. `originatingSourceLabel`: the label text used to describe the item's original source,
6. `certification`: if the dataset is signed, this property is used to contain the signature elements. This is an important element to prove the origin of a dataset as it is being sliced and diced during its evolution,
7. `swp:signature:` represents the signature of the dataset,
8. `swp:signatureMethod`: specifies the signature method,
9. `swp:authority`: defines the authority of the relationship between the item under provenance control and the dataset publisher,
10. `swp:valid-from` and `swp:valid-until`: these are the valid start and end dates of that (authority) relationship,
11. `processType`: specifies the type of transformation or conversion procedure carried out on the item's source, e.g. the transformation may be due to some scripts being run on the source data,
12. `prv:createdBy`: specifies the actor that executes an action on the item that is being recorded.
13. `prv:performedAt`: date when the transformation is done,
14. `prv:performedBy`: the URI of the actor that performs the recording of the provenance activity on the item.

6. EXPERIMENTS AND RESULTS

Our group, the EnAKTing group[5], is dedicated to solving fundamental problems in achieving an effective web of linked data, and as part of our work, we make use of some of the United Kingdom's government data. As part of our group's work, we recently converted a set of government data files from comma-separated-values (csv) to RDF datasets.

6.1 Source Datasets

Some of these data files were[6]:

- Mortality data:

  http://www.statistics.gov.uk/downloads/theme_population/Table_3_Deaths_Area_Local_Authority.xls,

[4]voidp's ontology is available at http://www.enakting.org/provenance/voidp/.

[5]http://enakting.org

[6]We advise that these csv files are checked for their contents.

- Population data: *http://www.statistics.gov.uk/downloads/theme_population/Mid-2003_Parl_Con_quinary_est.xls*,
- Energy:

  http://www.decc.gov.uk/assets/decc/statistics/regional/road_transport/file45728.xls,
- CO_2 emission:

  http://www.decc.gov.uk/assets/ decc/ statistics/climate _change/ 1_20100122174542 _e_@@ _localregionalco2 emissionsest20057.xls,
- Crime:

  www.homeoffice.gov.uk/rds/pdfs09/hosb1109chap7.xls.

A snippet of the RDF/Turtle schema representation for the Crime data[7] is shown below:

```
:TP2008_09 rdf:type crime:TimePeriod;
 dc:title "2001/02" ;
 scovo:min "2008-01-01"^^xsd:date;
 scovo:max "2009-12-31"^^xsd:date.
:Durham rdf:type crime:GeographicalRegion;
dc:title "Durham".
:Robbery rdf:type :CriminalOffenceType;
dc:title "Robbery".
:ds1_2_4 rdf:type scovo:Item;  rdf:value 170;
scovo:dataset :ds1; scovo:dimension :Robbery;
scovo:dimension :Durham;  scovo:dimension :TP2008_09.
```

We used voiD to describe these datasets. These datasets and their voiD descriptions were inserted into our RDF database, 4store[8]. The following snippet shows the RDF/Turtle schema representation of the voiD descriptions for one of the datasets, the crime dataset[9].

```
<http://crime.psi.enakting.org/id/void> a void:Dataset;
foaf:homepage <http://crime.psi.enakting.org/>;
rdfs:label "crime.psi.enakting.org Linked Data Repository";
dcterms:date "2010-09-14T16:54:31"^^xsd:date;
dcterms:title
  "crime.psi.enakting.org Linked Data Repository";
voidp:activity [ a voidp:Provenance;
voidp:item [ foaf:name
   <http://crime.psi.enakting.org/ds1>;
rdf:type  scovo:Dataset;
   rdfs:label
     "RECORDED CRIME STATISTICS 2008/09"@en ;
prv:createdBy [ rdf:type prv:Actor ;
   prv:performedBy <http://tomitola> ; ];
voidp:originatingSource [
voidp:originatingSourceURI
<http://www.homeoffice.gov.uk/rds/
   pdfs09/hosb1109chap7.xls> ;
voidp:originatingSourceLabel
   "Home Office UK"^^xsd:string; ];
voidp:processType
<http://void.rkbexplorer.com/id/dataset/
     d1d473f29a9091069644824242e9ae07> ;
to:hasBeginning
     "2010-09-14T16:54:31"^^xsd:dateTime ;
to:hasEnd
     "2010-09-14T20:54:31"^^xsd:dateTime ; ];
voidp:item [ foaf:name
   <http://crime.psi.enakting.org/id/Durham>;
rdf:type scovo:Dataset, prv:DataItem;
rdfs:label
   "Values of criminal offences for Durham
    for 2008/09."@en;
prv:createdBy [ rdf:type prv:Actor ;
   prv:performedAt
     "2010-09-14T16:54:31"^^xsd:dateTime ;
prv:performedBy <http://tomitola> ; ];
voidp:originatingSource [
voidp:originatingSourceURI
   <http://www.homeoffice.gov.uk/rds/pdfs09/
    hosb1109chap7.xls> ;
voidp:originatingSourceLabel
  "Home Office UK"^^xsd:string; ];];];]; .

<http://void.rkbexplorer.com/id/dataset/
    d1d473f29a9091069644824242e9ae07>
rdfs:label
"Data Transformation
  using a set of locally
  produced php scripts"^^xsd:string .
```

Figure 1 shows a typical usage scenario of our system.

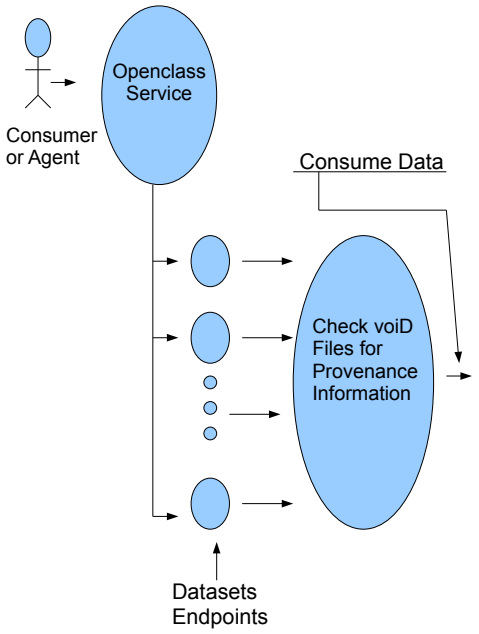

Figure 1: Typical System usage scenario

[7]The way we model the datasets were explained in [16].
[8]http://4store.org/
[9]The voiD descriptions used in this experiment can be found at http://152.78.189.49/voidp/. The provenance elements can be seen in the voiD descriptions.

Example Scenario Query.

An example query we are interested in is given below:

"Give the originating urls of the datasets for Robbery and female population for the County of Durham in the United Kingdom for 2004. Also give the CO_2 emission values and total energy consumption values for that same area. Only give datasets that are from the United Kingdom Home Office and from the United Kingdom's Department of Energy and Climate Change".

The Openclass Service.

OpenClass treats the federated schema of Linked Data as an augmentation of local data schemas that can be derived from the SPARQL endpoints regardless of cross-reference relationships in the data. The data schema is extracted from the triples of the local space that have compatible structures with RDF or OWL standards. These data schema can then be interrogated to find out their local SPARQL endpoints. In essence, Openclass acts as a directory service that, given some properties, checks its address space to find out which endpoints these properties are defined.

For example, to find out which endpoints have classes, instances, or dimensions that are labelled "Robbery", we made use of the SPARQL query below:

```
select ?subj  ?g where {
graph ?g {?subj
    <http://www.w3.org/2000/01/rdf-schema#label>
            "Robbery"
   }
};
```

These gave our (local) SPARQL endpoint which has the assertions for the dimension of "Robbery".

6.2 Provenance Queries

Once Openclass gives the SPARQL endpoints where these properties, dimensions, or classes are defined, we then query the respective endpoints to get their voiD descriptions. The queries expressed in our example query in section 6.1 are then run in a distributed fashion over the SPARQL endpoints.

The SPARQL query below queries our local crime endpoint for the Home Office dataset. Running this query gave us the originating source URL.

```
select distinct  ?o2  ?o3 ?o5 from
<http://crime.psi.enakting.org/id/void> where {
?s <http://purl.org/void/provenance/ns/activity> ?o1 .
     ?o1 ?p1 ?o2  .
     ?o2
<http://purl.org/void/provenance/ns/originatingSource>
     ?o3 .
     ?o3 voidp:originatingSourceLabel ?o4 .
     FILTER regex(?o4, ".*Office.*", "i") .
     ?o3 voidp:originatingSourceURI ?o5 .
};
```

We got the following result:

```
|?o2  |?o3 | ?o5 |
--------------------------------
| _:b1d841e00000000f7 |
_:b1d841e00000000f6  |
<http://www.homeoffice.gov.uk/rds/
pdfs09/hosb1109chap7.xls> |
```

Running the distributed query: "Give the originating urls of the datasets for Robbery and female population for the County of Durham in the United Kingdom for 2004. Also give the CO_2 emission values and total energy consumption values for that same area. Only give datasets that are from the United Kingdom Home Office and from the United Kingdom's Department of Energy and Climate Change", gave us the source urls that were stated in subsection 6.1 (Source Datasets).

7. CONCLUSIONS

In the open, often chaotic, world of the (Semantic) web, where diverse materials of disparate qualities are being made available, a mechanism that allows consumers to automatically find out the origin, i.e. provenance, of these materials is needed. In this paper, we describe **voidp**, a light-weight provenance extension for the voiD vocabulary that allows data publishers to add provenance metadata to the elements of their datasets. We enumerated its classes and properties, and described an experiment using a set of United Kingdom's public data to show how voidp can be utilised.

In future work, we will apply voidp to describe more datasets and will extend our results to be a foundation of a trust model. In addition, we will be using voidp as a basis of a semantic recommendation engine.

8. ACKNOWLEDGMENTS

This work was supported by the EnAKTing project, funded by EPSRC project number EP/G008493/1.

9. ADDITIONAL AUTHORS

10. REFERENCES

[1] K. Alexander, R. Cyganiak, M. Hausenblas, and J. Zhao. Describing linked datasets. *LDOW2009*, 2009.

[2] D. Artz and Y. Gil. A survey of trust in computer science and the semantic web. *Journal of Web Semantics*, 5(2), 2007.

[3] T. Berners-Lee. Cleaning up the user interface. *http://www.w3.org/DesignIssues/UI.html (retrieved Nov. 2010)*, 2009.

[4] C. Bizer. Semantic web publishing vocabulary (swp) user manual. *www4.wiwiss.fu-berlin.de/bizer/WIQA/swp/SWP-UserManual.pdf (retrieved Nov. 2010)*, 2006.

[5] C. Bizer. *Quality-Driven Information Filtering- In the Context of Web-Based Information Systems*. VDM Verlag, 2007.

[6] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. *International Conference on Database Theory, LNCS*, 1973, 2001.

[7] J. J. Carroll, C. Bizer, P. Hayes, and P. Stickler. Cleaning up the user interface. *WWW '05 Proceedings of the 14th international conference on World Wide Web*, 2005.

[8] C.Bizer, R. Cyganiak, and T. Heath. How to publish linked data on the wel. *http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/ (retrieved Nov. 2010)*.

[9] J. Cheney, L. Chiticariu, and W. C. Tan. Provenance in databases: Why, where and how. *Fundations and Trends in Databases*, 4(1), 2009.

[10] J. Golbeck. Trust on the world wide web: A survey. *Foundations and Trends in Web Science*, 1(2), 2008.

[11] O. Hartig. Provenance information in the web of data. *LDOW2009*, 2009.

[12] O. Hartig. Querying trust in rdf data with tsparql. *Lecture Notes in Computer Science*, 5554, 2009.

[13] O. Hartig and J. Zhao. Using web data provenance for quality assessment. *Proceedings of the 1st Int. Workshop on the Role of Semantic Web in Provenance Management, ISWC*, 2009.

[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] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gill, and e. a. P. Groth. The open provenance model core specification (v1.1). *Future Generation Computer Systems*, 2010.

[16] T. Omitola, C. L. Koumenides, I. O. Popov, Y. Yang, M. Salvadores, M. Szomzor, T. Berners-Lee, N. Gibbins, W. Hall, and N. S. m. c. schraefel. Put in your postcode, out comes the data: A case study. *ESWC*, 2010.

[17] J. Widom. Trio: A system for integrated management of data, accuracy, and lineage. *CIDR*, 2005.