# Provenance Information in the Web of Data

Olaf Hartig
Humboldt-Universität zu Berlin
Department of Computer Science
Berlin, Germany
hartig@informatik.hu-berlin.de

## ABSTRACT

The openness of the Web and the ease to combine linked data from different sources creates new challenges. Systems that consume linked data must evaluate quality and trustworthiness of the data. A common approach for data quality assessment is the analysis of provenance information. For this reason, this paper discusses provenance of data on the Web and proposes a suitable provenance model. While traditional provenance research usually addresses the creation of data, our provenance model also represents data access, a dimension of provenance that is particularly relevant in the context of Web data. Based on our model we identify options to obtain provenance information and we raise open questions concerning the publication of provenance-related metadata for linked data on the Web.

## Categories and Subject Descriptors

I.2.4 [**Computing Methodologies**]: Knowledge Representation Formalisms and Methods; H.3.3 [**Information Systems**]: Information Search and Retrieval

## Keywords

Provenance, Lineage, Web Data, Web of Data, Linked Data

## 1. INTRODUCTION

Today, a large amount of RDF data is published on the Web; large datasets are interlinked; new applications emerge that utilize this data in novel and innovative ways. An upcoming challenge that has to be addressed in these applications is the evaluation of qualities of the data retrieved from the Web, qualities such as accuracy, timeliness, reliability, and trustworthiness.

A recent study shows that one of the main factors that influence the trust of users in Web content is provenance [18]. Thus, a common approach for data quality assessment is the analysis of provenance information. “Information about provenance constitutes the proof of correctness [...] and [...] determines the quality and amount of trust [...]” [30]

Provenance information about a data item is information about the history of the item, starting from its creation, including information about its origins. Tan [30] distinguishes two granularities of provenance: workflow (or coarse-grained) provenance and data (or fine-grained) provenance. Workflow provenance represents “the entire history of the derivation of the final output of” [30] a workflow. Davidson et al. [14] provide an overview of provenance in workflow systems. Data provenance, in contrast, provides a more detailed view on the derivation of single pieces of data. A particular area of research on data provenance is provenance in databases which considers the provenance of query results. In this context, Buneman et al. [8] distinguish why- and where-provenance: why-provenance represents the origins that were involved in calculating a single entry of a query result; where-provenance refers to the exact locations an element of a query result entry has been extracted from. Green et al. [20] additionally introduce how-provenance that, in contrast to why-provenance, describes how the origins were involved in the calculation.

While a great many approaches exist that represent provenance [4, 29, 30], none of these explicitly addresses the characteristics of provenance of data from the Web. Web data provenance includes the access of data items on the Web, an information not required in the context of self-contained systems such as a DBMS or a workflow management system.

In this paper we discuss provenance of Web data. We aim to provide a base for research on the application of provenance information to assess qualities of linked data from the Web. Our main contributions are the following:

- We propose a provenance model that captures both, information about Web-based data access as well as information about the creation of data.
- We describe options to obtain provenance information by accessing metadata on the Web.
- We analyze vocabularies for RDF data that allow to describe provenance information.
- We identify open questions concerning the publication of provenance-related metadata for linked data on the Web.

This paper is structured as follows. First, Section 2 reviews related work. In Section 3 we introduce our provenance model for Web data. A discussion of options to obtain provenance information is given in Section 4. Finally, Section 5 raises open questions and Section 6 concludes this paper.

## 2. RELATED WORK

Representing and analyzing provenance is a topic of research since many years [4]. Simmhan et al. [29] provide a taxonomy of provenance characteristics. The authors differentiate between data-oriented approaches and process-oriented approaches. While data-oriented approaches focus

*LDOW2009*, April 20, 2009, Madrid, Spain.

on data items, process-oriented approaches emphasize information about the processes that consume and generate the data. Due to its level of abstraction our provenance model can be used as a basis for both types of approaches as well as for hybrid approaches.

An approach to model provenance on a more detailed level is the Open Provenance Model introduced by Moreau et al. [25]. Similar to our model, the authors distinguish three types of pieces of provenance information: artifacts, processes, and agents. The Open Provenance Model represents provenance by graphs. The nodes in these graph represent the artifacts, processes, and agents. The edges are directed and they have a predefined semantic depending on the type of the adjacent nodes. For instance, an edge that connects a process with an agent means the process was controlled by the agent. Some edges can be annotated with a use case-specific role. Due to its more detailed representation the Open Provenance Model can be used to realize the description of parts of a provenance graph that complies with our, more abstract model.

Bunemann et al. [7] raise several open questions for data provenance in the age of the Web. The authors identify three main issues: i) obtaining provenance information, ii) citing components of a digital library such as (components of) a document in another context, and iii) ensuring integrity of citations under the assumption that cited databases evolve. We address the first issue by discussing options to obtain provenance information in Section 4.1.

Harth et al. [21] argue for a provenance model for the Web that includes a "social dimension to associate provenance with the originator (typically a person) of a given piece of information." Given such a model it is possible to embed provenance-based quality assessments in the social context of users. We agree to the authors' request. With our provenance model we encourage to represent human actors and their relation to data items.

A more technical notion of provenance is represented by Ding et al. [15] who understand the provenance of RDF data as the RDF graphs of which parts of an analyzed RDF graph has been derived from. The authors argue that tracking complete RDF graphs is too coarse-grained and that a representation on the level of single RDF statements is unsuitable, too. For this reason, Ding et al. introduce RDF molecules as the finest sub-graphs that can be generated by a lossless decomposition of an RDF graph. Our provenance model represents data items on an abstract level. Thus, actual applications may use any level of granularity: RDF graphs, statements, or RDF molecules.

Hausenblas et al. [22] touch another aspect of Web data provenance. The authors distinguish sources of Web data based on the way these sources represent RDF data. Sources may contain RDF data in a non-serialized form (e.g. in-memory, in-store) or arbitrary data in a serialized form. Sources with serialized data may be i) RDF model-compliant and standalone, ii) RDF model-compliant and embedded, or iii) non-compliant to RDF model. Our provenance model also differentiates between the data itself and its representation in a document.

Another approach that considers provenance in the context of the Semantic Web has been developed in the Inference Web project. Da Silva et al. [12] describe a provenance infrastructure that supports "the extraction, maintenance and usage of knowledge provenance related to answers of web applications and services." The term knowledge provenance refers to information about the origin of knowledge and about the reasoning processes used to produce answers. In [11] the authors present the Proof Markup Language to describe justifications for results of an answering engine or a reasoner. A formal definition of justifications for entailments in OWL ontologies is provided by Horridge et al. [23]. These justifications may describe the execution of a specific kind of data creation processes represented by our provenance model.

## 3. A MODEL OF WEB DATA PROVENANCE

Provenance research in the context of databases [30] or in the context of workflows [14] usually focuses on the creation of data items such as query results and data products. In the majority of cases, these approaches apply a notion of the sources of a data item that is directly related to the creation process. To represent the provenance of data from the Web we need an additional dimension. Provenance information of Web data must comprise the aspect of publishing and accessing data on the Web. Questions such as who operates the service that provides a dataset are equally important as asking for the entity that created the data. For this reason, we suggest a provenance model for data from the Web that includes both dimensions, the creation and the access of data. In this section we describe our model: we introduce the basic elements, we present the data creation dimension, and we describe the representation of data access.

### 3.1 Basics of the Provenance Model

Provenance information can be used for various purposes. Possible uses are the estimation of data quality, the tracing of audit trails of data, the repetition of data derivations, the determination of liabilities, and the discovery of data [29]. The main purpose of our provenance model is to support the assessment of data qualities such as accuracy, reliability, and timeliness.

We propose to describe the provenance of a specific data item from the Web (e.g. a specific RDF graph or RDF statement) by a *provenance graph.* The nodes of provenance graphs are *provenance element*s that represent pieces of the provenance information about the data, pieces such as the actual creator of a specific dataset. Our provenance model identifies different types of provenance elements and it describes the relationships between these types and, thus, between the possible provenance elements in a provenance graph. Since provenance information for a data item may comprise information about source data a provenance graph for the data item may contain subgraphs that describe the provenance of the source data. Thus, our understanding of a *data item* does not only include RDF graphs and RDF statements but it also covers typical source data such as workflow results and database entries (Table 3 in Appendix B lists further examples for data items).

We broadly distinguish three types of provenance elements: actors, executions, and artifacts. An *actor* usually performs the *execution* of an action or a process which – in most cases – yields an *artifact* such as a specific dataset. An execution may include the use of artifacts which, in turn, might be the result of another execution. Furthermore, direct relationships between artifacts as well as between actors may exist. For instance, a specific company is responsible for its Web server.

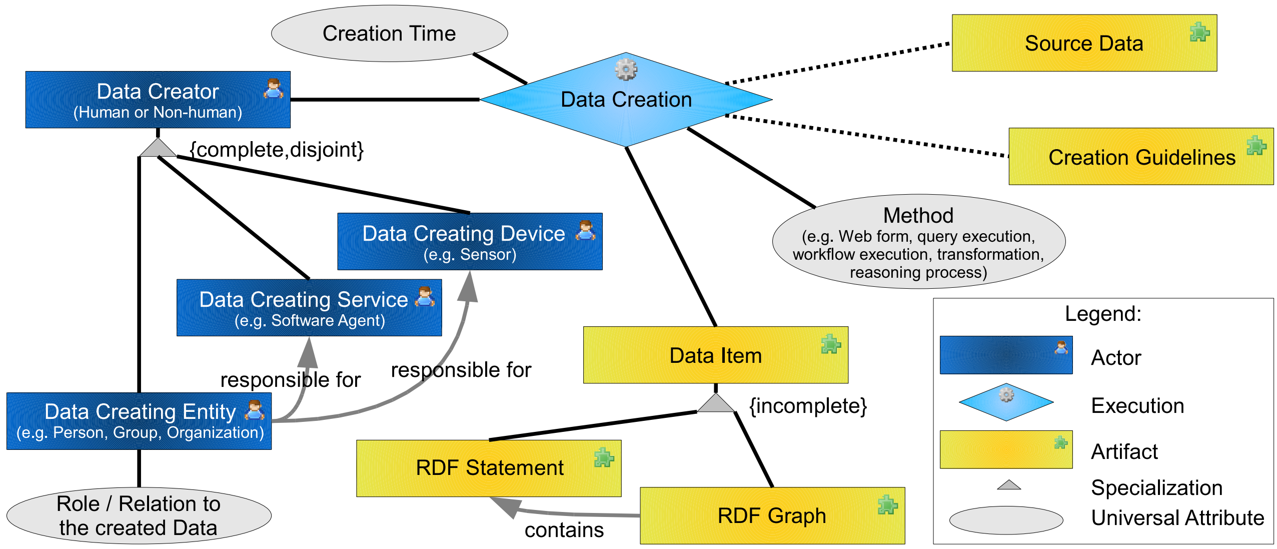

**Figure 1: Provenance information concerning the creation of data.**

Our model defines different *element types* as specializations of actors, executions, and artifacts. For example, data creators are a specific kind of actors and RDF graphs are a specific kind of artifacts. Each provenance element in a provenance graph corresponds to at least one of these types. The edges in a provenance graph correspond to the relationships between element types of the adjacent provenance elements.

All provenance elements have *attributes* that represent provenance-specific information about the elements. For instance, a specific data creator may have a name and may work for a well-known organization. The actual attributes and their extent depend on the element type and the needs in the application scenario. For some element types, however, all conceivable provenance elements will likely have attributes of the same kind. We call these attributes *universal attributes* and associate them with the corresponding element types in our model.

Please notice, to cover a broad variety of applications we designed our provenance model with generality in mind. Different applications of our model may have different needs with respect to the amount and granularity of the represented provenance information. For this reason, the elements of our model abstract from actual use cases; we do not suggest a specific implementation of provenance graphs nor do we prescribe attributes that must be used for specific provenance elements. However, we give examples for various provenance element types in Appendix B.

## 3.2 Data Creation

The creation of a data item can be a complex process such as executing a sophisticated workflow. Thus, provenance information about a data item may include a comprehensive description of the execution of data creation processes. On the other hand, the creation of data can be as straightforward as performing a simple action such as filling out a Web form. To represent this simple action as provenance of the derived data a few details may suffice. Furthermore, what is a simple action in one situation may be considered as a complex process in another case.

Our provenance model abstracts from the different notions of data creation. Figure 1 depicts the relationships of all element types that cover the data creation dimension. The central element type is the *data creation*. Data creations represent the execution of actions or processes that create new data items. Thus, in the provenance graph of a specific data item actual data creations are represented by provenance elements of the data creation type. All data creations have a *creation time* and use a *method* (cf. the universal attributes in Figure 1). Examples for data creation methods are the aforementioned completion of a Web form as well as the execution of a workflow, of a query, of a transformation, or of a reasoning process. The existence of further, provenance-related attributes of a data creation and the granularity of these attributes depend on the specific creation method and on the application of the provenance model. For instance, the execution of a query could be associated with a representation of how-provenance [20]; the inference of data could be represented with a justification [11][23].

Provenance elements that are associated with a data creation are the created data item, data creators, source data, and creation guidelines. *Data creators* are actors that perform the data creation. Our model distinguishes non-human and human data creators. Non-human creators are *data creating devices* such as sensors and *data creating services* such as software agents, reasoners, query engines, or workflow engines. Human data creators, called *data creating entities*, are persons, groups, organizations, etc. Data creating entities may create the data directly – as in the Web form example – or they are responsible for a non-human data creator that creates the data. A provenance-relevant attribute of all data creating entities is their *relation to the created data*. Further attributes depend on the actual provenance element and on the implementation of the provenance model. For example, a data creating service may implement a specific algorithm and, usually, has a version number and a developer.

A data creator often makes use of *source data* to create new data. Examples for source data are the content of a document used for machine learning, the statements in a knowledge base used to entail a new statement, and the entries in a database used to answer a query. The granularity by which a provenance graph represents source data depends on the use case. Notice, not every data creation uses source data as indicated by the dashed connection in Figure 1. However, all source data has provenance which should be represented as a subgraph in the provenance graph of the created data. Thus, a provenance graph may recursively contain subgraphs for source data, for the sources of source data, and so on.

Further input artifacts that may be associated with a data creation are the *creation guidelines* which guided the execution of the data creation. Examples for creation guidelines are transformation rules, mapping definitions, entailment rules, and database queries.

## 3.3 Data Access

A system that uses Web data must access this data from a provider on the Web. Information about this process and about the providers is important for a representation of provenance that aims to support the assessment of data qualities. Hence, our provenance model introduces element types that give attention to the data access dimension. Figure 2 depicts the main element types and their relationships.

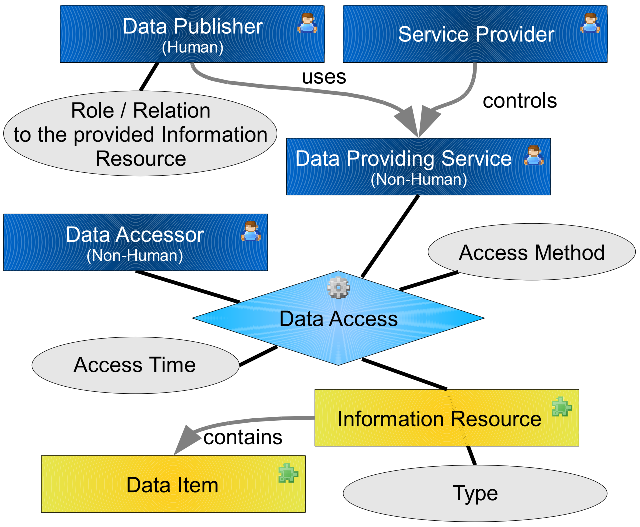

**Figure 2: Provenance information about data access on the Web.**

Data published on the Web is embedded in a host artifact, usually a document. Following the terminology of the W3C Technical Architecture Group we call this artifact an *information resource* [24]. Each information resource has a *type*, e.g., it is an RDF document or an HTML document.

A system, the *data accessor*, retrieves information resources from a provider. Our provenance model allows a detailed representation of providers by distinguishing data providing services, data publishers, and service providers. A *data providing service* is a non-human actor – usually a Web service or a server – that processes data access requests and actually sends the information resource over the Web. Provenance-related attributes of data providing services may be a description of the software that realizes the service. Notice, a data providing service in a provenance graph may also be a data creating service. For instance, a D2R server [1] creates RDF data from a relational database and provides this data as linked data in RDF documents, in HTML documents, or as the result of SPARQL queries.

*Data publishers* are persons, groups, or organizations that use a data providing service to publish data on the Web. Similar to the services, a data publisher may also be the data creating entity. The *service provider* element type represents entities such as a person, a groups, or an organization that controls a data providing service. Our model introduces this type because data publishers may publish their data on platforms that are provided by a third party, the service provider. Information about this third party may be relevant as provenance information. However, a data publisher may administer its own service and, thus, could be the service provider itself.

Our provenance model represents the actual execution of a data access by an element type called *data access*. Provenance information that is common to all provenance elements of this type is the *access time* and the *access method*. A major access method an HTTP-based resource request where the URI of the requested resource would be another provenance-related attribute of the corresponding provenance elements. Additional attributes of data access provenance elements may represent the content negotiation [17, Section 12] and possible redirections [17, Section 10.3] that happened during the data access. Further examples for access methods are API-based data access and its specialization, query-based access. Provenance information in these cases are the API call with its parameters and the issued query, respectively. Notice, query-based data access usually is a data creation too.

Further provenance information not considered so far is the availability and validity of digital signatures. Since this information is important to assess the quality of data our provenance model contains additional element types, namely intergrity assurances, digital signatures, public keys and signers (cf. Figure 3).

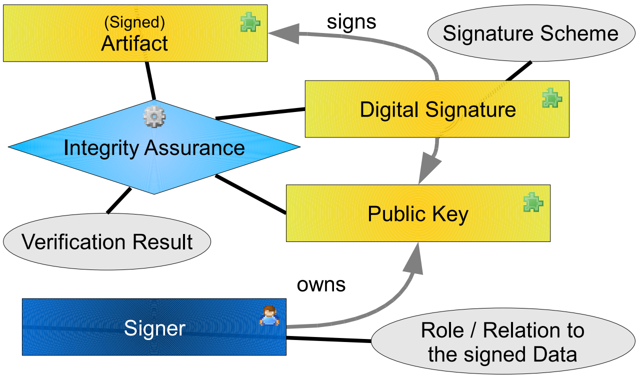

**Figure 3: Signature verifiability is a part of Web data provenance.**

An *intergrity assurance* basically represents the verification of a *digital signature* for the signed artifact. The verification requires the *public key* of the *signer*. Intergrity assurances are associated with information about the *result* of the verification. Digital signatures have several properties that are relevant as provenance information, e.g., the date of issue and the *signature scheme* [19]. The provenance element for a public key could describe the creation and the expiration date.

# 4. OBTAINING PROVENANCE INFORMATION

A system that applies our provenance model generates provenance graphs for data items. To create the provenance elements of such a graph the system has to collect different pieces of provenance information automatically. In this section we discuss options to obtain provenance information for Web data.

Some pieces of provenance information can be recorded by a system; for other pieces the system relies on metadata provided by third parties. Thus, we distinguish recordable provenance information and metadata-reliant provenance information. Basically, *recordable provenance information* is information on executions that are performed by the system itself or that can sufficiently be monitored by the system. Usually, these executions are data accesses initiated by the system, signature verifications, and local data creations. *Metadata-reliant provenance information*, in contrast, can not be recorded automatically but requires the evaluation of metadata that is published on the Web. Metadata-reliant provenance information comprises information about executions inaccessible to the system as well as information about actors and artifacts involved in these executions. Furthermore, obtaining more exhaustive provenance information about certain actors involved in accessible executions may also require metadata. For instance, even if a system can record information about a self-initiated data access a proper representation of the involved providers requires metadata.

Recording provenance information is a fundamental topic of provenance research. For instance, Bose and Frew [4] study scientific workflow management systems that aim to track provenance for data products; Horridge et al. [23] present concepts to compute justifications for entailments in ontologies; Tan [30] discusses different approaches to propagate and to compute provenance of query results in database systems. The concepts developed in these contexts can be adapted to generate recordable provenance information in the context of Web data. On that account, we focus on metadata-reliant provenance information in the remainder of this section. We identify methods to access relevant metadata on the Web, we analyze vocabularies that allow a representation of provenance-related metadata, and we study the existence of such metadata on the Web.

## 4.1 Accessing Provenance-Related Metadata

Provenance-relevant metadata is either directly attached to a data item or its host document or it is available as additional data on the Web. Examples for attached metadata are RDF statements about an RDF graph that contains the statements, author and creation date of blog entries added to a syndication feed, or information about an image embedded in the image file. Both, attached metadata and detached metadata, may be represented in RDF using vocabularies as described in Section 4.2 or it may be data of another form. In the following, we present options to discover detached metadata on the Web.

Accessing data on the Web is often based on HTTP URIs. Since these URIs are grounded in the Domain Name System (DNS) it is possible to query a WHOIS [13] service in order to get provenance information about the accessed data item. However, the responses of WHOIS services are hardly usable for automatic evaluation because the WHOIS protocol does not prescribe a standard form to structure returned data.

A source of data about the content provided by a Web server are sitemaps. A sitemap is an XML document that informs search engine crawlers about URLs on a website. The semantic sitemap approach [10] extends these documents with information about the location of RDF data and about alternative means to access this data (e.g. data dumps and SPARQL endpoints). Even if the information in a semantic sitemap is marginally provenance-related an important element is the specification of a URI that represents referenced datasets. Given the provider follows the linked data principles a look-up of these URIs will yield RDF-based metadata that may describe provenance information about the datasets.

Using linked data, in any case, is an important approach to discover provenance-related metadata. As with the URI of an RDF dataset it is possible to look up any HTTP URI that represents a piece of provenance (e.g. the URI of a data item such as a named RDF graph [9] or the URI of an actor). Moreover, collecting provenance information may involve following RDF links in order to get more complete information during the generation of provenance graphs.

Another method to discover metadata about Web resources is POWDER [28], the Protocol for Web Description Resources. POWDER introduces so called description resources to describe resources on the Web. These descriptions are either based on RDF data or on simple keywords (i.e. tags) and they may contain provenance information.

A specific kind of actors represented by our provenance model are Web services that create or provide data. Different standards exist to describe Web services [32]. These descriptions may also contain information that are relevant for provenance graphs.

## 4.2 Provenance-Related Vocabularies

Various vocabularies exist that allow to describe provenance information with RDF data. In the following, we describe these vocabularies and we relate the classes and properties defined by these vocabularies to the elements of our provenance model. Afterwards, we study the presence of the vocabularies in the Web.

A popular standard to represent general-purpose metadata are the Dublin Core Metadata Terms [16] which are available as an RDFS Schema. The following properties defined by this schema can be associated with a resource to describe provenance information:

- `dcterms:contributor`[1], `dcterms:creator` – The contributor of a resource is defined as "an entity responsible for making contributions to the resource" [16] and the creator is "an entity primarily responsible for making the resource." [16] Thus, these properties may be used to obtain information about data creators of a data item. However, the actual type of the referenced

[1] We use the following namespace prefixes in this paper:
dcterms: http://purl.org/dc/terms/
dc11: http://purl.org/dc/elements/1.1/
foaf: http://xmlns.com/foaf/0.1/
sioc: http://rdfs.org/sioc/ns#
swp: http://www.w3.org/2004/03/trix/swp-2/
wot: http://xmlns.com/wot/0.1/
iwProv: http://inferenceweb.stanford.edu/2006/06/pml-provenance.owl#
ouzo: http://www.mygrid.org.uk/provenance#
cs: http://purl.org/vocab/changeset/schema#

creators as well as their role in the data creation process remain unclear because the Dublin Core schema does not distinguish different types of data creators as our provenance model does. Analyzing data about the creator (or about the contributor) may yield further information (e.g. the type) that can help to derive a more precise provenance graph.

- `dcterms:source` – The source of a resource is "a related resource from which the described resource is derived." [16] With this property it is possible to create provenance elements associated as source data with a data creation element.

- `dcterms:created` – This property specifies the creation date of a resource and can be used to set the creation time attribute associated with the execution of a data creation.

- `dcterms:modified` – This property specifies the date on which a resource has been changed. We propose to represent the modification of a data item as a data creation which creates a new, modified version of the original data item. The creation time attribute associated with this data creation can be set using the `dcterms:modified` property.

- `dcterms:publisher` – The publisher of a resource is "an entity responsible for making the resource available" [16] This property may be used to obtain information about a provider of an information resource whereas the actual type of the provider (data providing service, data publisher, or service provider) remains unclear.

- `dcterms:provenance` – This property links a resource to "a statement of any changes in ownership and custody of the resource since its creation that are significant for its authenticity, integrity, and interpretation." [16] Due to the very general definition, it is difficult to use such a provenance statement during the creation of a provenance graph.

The Friend of a Friend (FOAF) vocabulary [6] provides classes and properties to describe entities such as persons, organizations, groups and software agents. FOAF-based descriptions can be used to obtain basic information about actors (e.g. names, group membership, email addresses, identifying online accounts). Furthermore, FOAF contains the property `foaf:maker` and its inverse `foaf:made` to relate the described entities to resources made by the entities. These properties can be used to identify the data creator of a data item. However, using these properties raises the same questions as for the `dcterms:creator` property.

The Semantically-Interlinked Online Communities (SIOC) ontology [2] describes information from online communities. The ontology associates SIOC items such as blog posts, comments, and e-mail messages to users that are identified by their online accounts. The following properties describe provenance-relevant information:

- `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. In contrast to the corresponding Dublin Core and FOAF properties, representing the referenced users according to our provenance model is less ambiguous: creators and modifiers referenced in a SIOC-based description are data creating entities.

- `sioc:has_owner`, `sioc:owner_of` – These properties express ownership of SIOC items. This information may provide an indication of the relation of a data publisher to provided data and, thus, 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 versions of a SIOC item with each other and could be used to create relationships between artifacts in a provenance graph.

The Semantic Web Publishing Vocabulary (SWP) [9] enables the description of information about provision of data. With SWP it is possible to represent the attitude of a legal person to an RDF graph. SWP supports two attitudes: claiming the graph is true and quoting the graph without a comment on its truth. These commitments towards the truth can be used to derive a data publisher's or a data creating entity's relation to provided or created artifacts. Furthermore, the SWP allows to describe digests and digital signatures of RDF graphs and to represent public keys. Similarly, the Web Of Trust schema (WOT) [5] enables descriptions that document the use of public key cryptography tools to sign documents. However, two differences between WOT and the signature-related part of SWP exist. First, digital signatures in SWP are represented as RDF data; WOT, in contrast, refers to signatures that are individually encoded in dedicated documents. Second, while the digital signatures described with WOT sign information resources, the signatures in SWP-based descriptions sign a specific kind of data item, namely RDF graphs. However, both, SWP-based descriptions as well as WOT-based descriptions, can be used to obtain information about public key and digital signatures in order to represent them in a provenance graph.

Further vocabularies that can be used to describe provenance information of specific types of data items are the following:

- The Ontology Metadata Vocabulary (OMV) [27] describes ontologies. OMV includes properties for creators, contributers, reviewers, and creation and modification dates.

- The Proof Markup Language [11] describes justifications for results of an answering engine or an inference process.

- The Changeset Vocabulary [31] describes changes to RDF-based resource descriptions.

- The Ouzo Provenance Ontology [33] describes the run of a (scientific) workflow, the processed data, and the entities responsible for the workflow run.

## 4.3 Existence of Provenance Metadata

To study the existence of metadata on the Web that uses the aforementioned vocabularies and their provenance-relevant properties we utilized two Web data indexes available on the Web, namely the Web service Ping the Semantic Web (PTSW) [3] and the Sindice search engine [26]. PTSW is a service that receives notifications from different applications that create, update, or discover RDF documents on the Web; PTSW aggregates these notifications and provides an up-to-date list of existing RDF documents. Furthermore, the PTSW website provides different statistics. Currently, PTSW indexes 1,073,218[2] RDF documents. For each of the vocabularies discussed before Table 1 presents the number of documents registered at PTSW that use this vocabulary. As the numbers indicate, FOAF and SIOC are widely used. The other vocabularies are not at all or they are used in an insignificant number of documents.

**Table 1: Number of RDF documents known to PTSW that use a vocabulary (as of Feb. 7, 2009).**

| vocabulary | occurence |
|---|---:|
| Dublin Core Metadata Terms | 121 |
| Dublin Core Metadata Element (legacy) | 9 |
| FOAF | 989,263 |
| SIOC | 127,974 |
| SWP | 1 |
| WOT | 101 |
| Proof Markup Language | 0 |
| Ouzo Provenance Ontology | 0 |
| Changeset Vocabulary | 0 |

The statistics of PTSW may not be representative because they heavily depend on the applications that notify PTSW. For this reason, we utilized the Sindice search engine for another inquiry. Sindice indexes structured data from the Web. We queried Sindice for documents that contain RDF statements with the provenance-relevant properties of the vocabularies. Table 2 in Appendix A lists the number of documents in the Sindice index for each property. According to Sindice, Dublin Core is roughly as often used as FOAF,a conclusion that cannot be drawn from the PTSW statistics. Furthermore, in contrast to the PTSW numbers, the Sindice queries reveal that the legacy Dublin Core metadata elements are used more widely than the recommended new metadata terms. Consistent with the findings discovered in the PTSW statistics, the SWP, the Proof Markup Language, the Ouzo Provenance Ontology, and the Changeset Vocabulary are not used at all[3]. Moreover, considering that Sindice currently indexes about 48.99 million documents the numbers for the other vocabularies are not satisfying either. Thus, we conclude that there is only very little provenance-related, RDF-based metadata available on the Web.

## 5. OPEN QUESTIONS

Our analysis of vocabularies that allow to express provenance information reveals two problems. First, the vocabularies are partly unsuitable and lack certain features. Our review of the vocabularies shows a lack of unambiguousness for certain properties. In particular, it is difficult to distinguish between the different types of providers and it is impossible to express the actual relationships between providers of different types. The same holds for data creators. Furthermore, it is impossible to describe the execution of a data access. These descriptions may be required to document the access of source data executed by a third party.

To overcome these issues we propose to develop a vocabulary that enables data publishers to describe the provenance of the provided data more precisely. This new vocabulary may refine existing vocabularies. Nonetheless, the development of this new vocabulary should be based on the presented provenance model.

The second problem is the general lack of provenance-related metadata in the Web of linked data. Reasons might be the lack of suitable vocabularies, a lack of usable tools to generate provenance-related metadata, and ignorance or at least a lack of sensitization. The first two reasons can be ascribed to technical problems that should be solvable by the development of the proposed vocabulary and by the implementation of corresponding tools. The lack of sensitization is a more general problem that must be addressed by the linked data community. A possible approach may evolve based on the recently released Vocabulary of Interlinked Datasets (voiD) [34] which is a vocabulary to describe the content of RDF-based datasets and the links between different datasets. Since voiD enables the discovery and usage of linked datasets it may raise the awareness of publishers to provide metadata for their datasets. This understanding may be used to motivate publishers to provide provenance information along with their voiD descriptions.

## 6. CONCLUSIONS

In this paper we propose a provenance model for Web data and we discuss options to obtain provenance information. In contrast to provenance research in areas such as workflows and databases the analysis of the Web data provenance must include information about the access of data in the Web. Thus, our provenance model includes two dimensions: data creation and data access. By specifying the relationships of rather general types of pieces of provenance information our model describes provenance on an abstract level. This generality gives applications a choice to refine the model according to their use case.

Based on our provenance model we describe options to obtain provenance information and we analyze vocabularies to express such information. Our analysis identifies several open questions. We aim to address these questions in the future.

As further future work we will develop concepts to estimate the trustworthiness of Web data based on our provenance model. These estimations have to consider subjective assessments of the elements in a provenance graph (e.g. the reliability of a data creator) as well as the trustworthiness of the data that has been used to create the provenance graph.

## 7. ACKNOWLEDGMENT

We thank Jun Zhao for her valuable feedback on our provenance model.

## 8. REFERENCES

[1] C. Bizer and R. Cyganiak. D2R Server – Publishing Relational Databases on the Semantic Web. Poster at the 5th International Semantic Web Conference (ISWC), Nov. 2006.

[2] U. Bojars and J. G. Breslin. SIOC Core Ontology Specification, Revision 1.30. Online, Jan. 2009. Retrieved Feb. 7.

[2] All numbers are from February 7, 2009.

[3] Occurences of 1 to 3 in Table 2 refer to documents that specify the corresponding properties.

[3] U. Bojars, A. Passant, F. Giasson, and J. Breslin. An Architecture to Discover and Query Decentralized RDF Data. In *Proceedings of the 3rd Workshop on Scripting for the Semantic Web (SFSW) at ESWC*, June 2007.

[4] R. Bose and J. Frew. Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys*, 37(1):1–28, Mar. 2005.

[5] D. Brickley. Web Of Trust RDF Ontology. Online, Feb. 2004. Retrieved Feb. 7.

[6] D. Brickley and L. Miller. FOAF Vocabulary Specification. Online, Nov. 2007. Retrieved Feb. 7.

[7] 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)*. Springer, Dec. 2000.

[8] P. Buneman, S. Khanna, and W. C. Tan. Why and Where: A Characterization of Data Provenance. In *Proceedings of the 8th International Conference on Database Theory (ICDT)*. Springer, Jan. 2001.

[9] 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 (WWW )*. ACM Press, May 2005.

[10] R. Cyganiak, H. Stenzhorn, R. Delbru, S. Decker, and G. Tummarello. Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web. In *Proceedings of the 5th European Semantic Web Conference (ESWC)*. Springer, June 2008.

[11] P. P. da Silva, D. L. McGuinness, and R. Fikes. A Proof Markup Language for Semantic Web Services. *Information Systems*, 31(4-5):381–395, June 2006.

[12] P. P. da Silva, D. L. McGuinness, and R. McCool. Knowledge Provenance Infrastructure. *Data Engineering Bulletin*, 26(4):26–32, Dec. 2003.

[13] L. Daigle. WHOIS Protocol Specification. IETF RFC 3912, Sept. 2004.

[14] S. B. Davidson, S. C. Boulakia, A. Eyal, B. Ludäscher, T. M. McPhillips, S. Bowers, M. K. Anand, and J. Freire. Provenance in Scientific Workflow Systems. *IEEE Data Engineering Bulletin*, 30(4):44–50, Dec. 2007.

[15] L. Ding, T. Finin, Y. Peng, P. P. da Silva, and D. L. McGuinness. Tracking RDF Graph Provenance using RDF Molecules. Technical Report TR-CS-05-06, UMBC, Apr. 2005.

[16] Dublin Core Metadata Initiative Usage Board. DCMI Metadata Terms. Online, Jan. 2008. Retrieved Feb. 7.

[17] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, and T. Berners-Lee. Hypertext Transfer Protocol – HTTP/1.1. IETF RFC 2068, Jan. 1997.

[18] Y. Gil and D. Artz. Towards Content Trust of Web Resources. *Journal of Web Semantics*, 5(4):227–239, Dec. 2007.

[19] S. Goldwasser, S. Micali, and R. Rivest. A Digital Signature Scheme Secure Against Adaptive Chosen-Message Attacks. *SIAM Journal on Computing*, 17(2):281–308, Apr. 1988.

[20] T. J. Green, G. Karvounarakis, and V. Tannen. Provenance Semirings. In *Proceedings of the 26th Symposium on Principles of Database Systems (PODS)*. ACM, June 2007.

[21] A. Harth, A. Polleres, and S. Decker. Towards a Social Provenance Model for the Web. In *Proceedings of the Workshop on Principles of Provenance*, Nov. 2007.

[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.

[23] M. Horridge, B. Parsia, and U. Sattler. Laconic and Precise Justifications in OWL. In *Proceedings of the 7th International Semantic Web Conference (ISWC)*. Springer, Oct. 2008.

[24] I. Jacobs and N. Walsh. Architecture of the World Wide Web, Volume One. W3C Recommendation, Online, Dec. 2004. Retrieved Feb. 7.

[25] L. Moreau, B. Plale, S. Miles, C. Goble, P. Missier, R. Barga, Y. Simmhan, J. Futrelle, R. McGrath, J. Myers, P. Paulson, S. Bowers, B. Ludaescher, N. Kwasnikowska, J. Van den Bussche, T. Ellkvist, J. Freire, and P. Groth. The Open Provenance Model. Technical report, Electronics and Computer Science, University of Southampton, 2008.

[26] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: A Document-oriented Lookup Index for Open Linked Data. *International Journal of Metadata, Semantics and Ontologies*, 3(1):37–52, 2008.

[27] R. Palma, J. Hartmann, and P. Haase. OMV Ontology Metadata Vocabulary for the Semantic Web, v2.4. Online, Jan. 2008. Retrieved Feb. 7.

[28] K. Scheppe. Protocol for Web Description Resources (POWDER): Primer. W3C Working Draft, Online, Nov. 2008. Retrieved Feb. 7.

[29] Y. Simmhan, B. Plale, and D. Gannon. A Survey of Data Provenance in e-Science. *SIGMOD Record*, 34(3):31–36, Sept. 2005.

[30] W. C. Tan. Provenance in Databases: Past, Current, and Future. *IEEE Data Engineering Bulletin*, 30(4):3–12, Dec. 2007.

[31] S. Tunnicliffe and I. Davis. Changeset Vocabulary. Online, Mar. 2006. Retrieved Feb. 7.

[32] K. Verma and A. Sheth. Semantically Annotating a Web Service. *IEEE Internet Computing*, 11(2):83–85, Mar. 2007.

[33] J. Zhao. *A conceptual model for e-science provenance*. PhD thesis, University of Manchester, June 2007.

[34] J. Zhao, K. Alexander, M. Hausenblas, and R. Cyganiak. Vocabulary of Interlinked Datasets. Online, Jan. 2009. Retrieved Feb. 7.

# APPENDIX

## A. USAGE OF VOCABULARIES

This appendix details the result of a Sindice-based search for documents that contain certain provenance-relevant properties. Table 2 lists each property together with the number of documents that contain at least one RDF statement with the property. The numbers have been recorded on February 7, 2009, by querying the Sindice search engine.

**Table 2: Provenance-relevant properties and the number of documents in which they occur at least once (according to the Sindice search engine).**

| property | occurences |
|---|---:|
| dcterms:creator | 134 |
| dc11:creator | about 24,150 |
| dcterms:contributor | 11 |
| dc11:contributor | 465 |
| dcterms:source | 1 |
| dc11:source | about 3,630 |
| dcterms:created | about 73,010 |
| dc11:created | about 9,710 |
| dcterms:modified | about 2,320 |
| dc11:modified | about 9,700 |
| dcterms:publisher | 87 |
| dc11:publisher | 808 |
| dcterms:provenance | 7 |
| foaf:made | about 5,420 |
| foaf:maker | about 29,370 |
| sioc:creator_of | about 1,370 |
| sioc:has_creator | about 4,520 |
| sioc:modifier_of | 3 |
| sioc:has_modifier | 4 |
| sioc:owner_of | about 3,020 |
| sioc:has_owner | about 553 |
| sioc:earlier_version | 0 |
| sioc:later_version | 0 |
| sioc:next_version | 3 |
| sioc:previous_version | 3 |
| swp:assertedBy | 0 |
| swp:authority | 0 |
| swp:quotedBy | 0 |
| swp:validUntil | 0 |
| wot:assurance | 135 |
| wot:fingerprint | 52 |
| wot:hasKey | 23 |
| wot:hex_id | 48 |
| wot:identity | 36 |
| wot:length | 43 |
| wot:pubkeyAddress | 54 |
| wot:sigdate | 8 |
| wot:signed | 3 |
| wot:signer | 2 |
| iwProv:hasMember | 1 |
| iwProv:isMemberOf | 1 |
| iwProv:hasPublisher | 1 |
| iwProv:hasPublicationDateTime | 1 |
| iwProv:hasUsageDateTime | 1 |
| iwProv:hasSource | 1 |
| iwProv:hasInferenceEngineRule | 1 |
| iwProv:usesInferenceEngine | 1 |
| ouzo:belongsTo | 2 |
| ouzo:dataDerivedFrom | 2 |
| ouzo:launchedBy | 2 |
| ouzo:processInput | 2 |
| ouzo:runsWorkflow | 2 |
| cs:createdDate | 3 |
| cs:creatorName | 3 |

# B. PROVENANCE ITEM TYPES

This appendix lists examples for various provenance element types in the data creation dimension (cf. Table 3) and in the data access dimension (cf. Table 4).

**Table 3: Examples of provenance element types in the data creation dimension.**

| element type | examples |
|---|---|
| data item | RDF statement |
| | RDF graph |
| | subgraph of an RDF graph |
| | axiom in a knowledge base |
| | data product created by workflow |
| | set of data products |
| | result of a query |
| | table in a database |
| | tuple in a database table |
| data creation | completion of a Web form |
| | execution of a workflow |
| | execution of a transformation |
| | automatic reasoning |
| | execution of a database query |
| | execution of a search query |
| | mapping from other data models |
| | machine learning |
| data creating entity | person, group, organization |
| data creating device | sensor |
| data creating service | software agent |
| | workflow engine |
| | reasoner |
| | query engine |
| | search index |
| | data wrapper (e.g. D2R Server) |
| | (batch) script interpreter |
| source data | content of a document |
| | (statements in) a dataset |
| | (data in) a database |
| creation guidelines | workflow model |
| | transformation rules |
| | entailment/inference rules |
| | database query |
| | search query |
| | mapping definitions |

**Table 4: Examples of provenance element types in the data access dimension.**

| element type | examples |
|---|---|
| data access | resource-based data access |
| | API-based data access |
| | query-based data access |
| data providing service | Web server |
| | Web service |
| | Web-based query interface |
| data publisher | person, group, organization |
| service provider | person, group, organization |