Using Web Data Provenance for Quality Assessment

Olaf Hartig
Databases and Information Systems Research Group
Department of Computer Science
Humboldt-Universität zu Berlin
Email: hartig@informatik.hu-berlin.de

Jun Zhao
Image Bioinformatics Research Group
Department of Zoology
University of Oxford
Email: jun.zhao@zoo.ox.ac.uk

***Abstract*—The Web of Data cannot be a trustworthy data source unless an approach for evaluating the quality of data on the Web is established and integrated as part of the data publication and access process. In this paper, we propose an approach of using provenance information about the data on the Web to assess their quality and trustworthiness. Our contributions include a model for Web data provenance and an assessment method that can be adapted for specific quality criteria. We demonstrate how this method can be used to evaluate the timeliness of data on the Web, to reflect how up-to-date the data is. We also propose a possible solution to deal with missing provenance information by associating certainty values with calculated quality values.**

I. Introduction

With the growth of the open-accessible Web of Data [1] the needs for evaluating the quality of the data in applications are becoming more and more pressing. Information quality research has been successfully applied to evaluate the quality of organizational information and to monitor the improvement of work practice [2]. Quality assessment of data on Web should be a paramount task in order to ensure that the most appropriate and trustworthy data are made available and delivered to the users. Scientific applications built upon the Web of Data will be of little value if scientists are skeptical of the quality of data; financial systems will be untrustworthy and fragile without any policies for quality control and evaluation.

To assess the quality of data, we need to identify the types of information that can be used for evaluation and a method for calculating quality values. In this paper, we present an approach that uses provenance information to assess quality of data on the Web; and we propose a generic assessment procedure that can be adapted for evaluating specific quality-criteria, such as accuracy and timeliness.

As the base of our approach we introduce a provenance model tailored to the needs for tracking and tracing provenance information about data on the Web. This model not only represents the creation of a data item, but also describes provenance information about the entities who made the data accessible on the Web [3]. We call this *Web data provenance*.

Many existing information quality assessment approaches are based on information contributed by users. In this paper, we focus on using a quantitative approach for calculating quality of data. This assessment approach takes three steps: collecting the elements of provenance information needed for quality assessment, then deciding on the influence of these elements on the assessment, and, finally, applying a function to calculate the quality.

To demonstrate how design decisions can be made when developing this method into assessing specific quality criteria we walk through the development using the *timeliness* criterion as an exemplar. Since the provenance information required for quality assessment might be incomplete or fragmentary, the assessment method must be capable to deal with missing information. We introduce a possible approach of associating certainty values with the calculated timeliness value.

This paper is structured as follows. Section II reviews related work and Section III introduces the model for Web data provenance. Section IV describes our assessment method that can be adapted for specific quality criterion, like timeliness, as demonstrated in Section V. We conclude in Section VI.

II. Related Work

In this paper we consider information quality (IQ) as "an aggregated value of multiple IQ-criteria" [4], such as accuracy, completeness, believability, and timeliness. The assessment of information quality can be regarded as "the process of assigning numerical values (IQ-scores) to IQ-criteria" [4]. IQ assessment is known to be hard [4]. Although there are many related work on conceptualizing the problem of IQ and its assessment, there are fewer work proposing concrete methods for quantifying the quality assessment. In the following we first introduce different approaches for IQ assessment in general, and then we focus on provenance-based.

Lee et al. [2] propose a quality assessment methodology that measures IQ from four quadrants: soundness, dependability, usefulness, and usable information. Each dimension includes several IQ-criteria. For example the dependability of information is measured by its timeliness and security. A questionnaire is designed to measure users' feedback to each IQ-criterion in a scale of 0-10. The assessment value for each quadrant is computed as a mean of the measurements of its constituent IQ-criteria. Similarly, Bobrowski et al. [5] also use questionnaires to assess information quality. Both methods, although being quantitative, are based on subjective, users' inputs.

In certain circumstances, an automatic assessment of information quality could be feasible with available metadata information and reliable auto-assessment techniques. Depending on the application and users' needs, this automatic approach could be more desirable than a subjective, manual approach.

Motro and Rakov propose automated assessment methods for evaluating the soundness and completeness of data sources [6]. Gruser et al. present a prediction algorithm to learn and predict response times of Web data sources [7]. Ballou et al. [8] introduce a quantitative assessment method for measuring and calculating the timeliness of data. Their formulas laid the foundation for our work and will be detailed in Section V.

Ballou et al.'s method for assessing the timeliness is partially based on the provenance information of a data item, e.g., the time when the data was obtained. Provenance metadata has been used to evaluate other IQ-criteria, such as the believability and trustworthiness. Wong et al. [9] use information about the types of services or data involved in a data creation process to validate the believability of derived data items. Golbeck and Mannes [10] use provenance of user-contributed annotations to compute trust values and to recommend how much a user should trust others. This method does not compute the trustworthiness of the annotations themselves using provenance.

III. A Model for Web Data Provenance

Our provenance-based IQ assessment method is based on our model for Web data provenance. We give a brief introduction to the model in this section. A detailed discussion of the model can be found in [3].

Traditional provenance research usually addresses the creation of data. While many approaches exist that represent provenance [11], [12], none of these explicitly addresses the characteristics of data that was not only created but also retrieved over the Web. Provenance of data from the Web comprises information about the entities that published the data and that made it accessible on the Web, information not required in the context of self-contained systems such as a DBMS or a workflow management system. Hence, our model for Web data provenance comprises two dimensions: data creation and data access.

Our model identifies types of so called *provenance elements* and the relationships between these types. The provenance elements represent pieces of provenance information; such an element might be the actual creator of a specific data item what makes this element an instance of the 'data creator' type. The types are classified in three categories: actors, executions, and artifacts. An *actor* usually performed the *execution* of an action or a process which – in most cases – yielded an *artifact* such as a specific data item. An execution might have included 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 was responsible for its Web server. All other element types are specializations of actors, executions, and artifacts.

The central type in the data creation dimension is the *data creation* execution by which a data item was created. A data creation was performed by a *data creator*. For the creation *source data* and *creation guidelines* could have been used by the data creator.

The data access dimension centers around *data access* executions. *Data accessors* perform data access executions to retrieve data items contained in *documents* from a provider on the Web. To enable a detailed representation of providers the model distinguishes *data providing services* that process data access requests and send the documents over the Web, *data publishers* who use data providing services to publish their data, and *service providers* who operate data providing services. Furthermore, the model represents the execution of *integrity verifications* of artifacts and the results thereof.

Based on the element types and their relationships identified by our provenance model it is possible to represent provenance of data items from the Web by, so called, *provenance graph*s. The nodes in these graphs are the provenance elements; the edges correspond to the relationships between the element types of adjacent elements; edges are labeled with the relationship name. Notice, to allow for a wide range of applications of our model we do not prescribe a specific granularity by which provenance information has to be described in provenance graphs. For instance, a data item could be a whole linked dataset as well as a single RDF statement, depending on the granularity required for the use case at hand. A data item could have been created by the use of creation guidelines and source data which also have provenance. This provenance should be represented by subgraphs in the provenance graph of the created data item.

Formally, we represent a provenance graph as a tuple $(PE, R, type, attr)$ where

- PE denotes the set of provenance elements in the graph,
- $R \subseteq PE \times PE \times RN$ denotes the labeled edges in the graph where RN is the set of relationship names as introduced by our provenance model,
- $type : PE \rightarrow \mathfrak{P}(T)$ is a mapping that associates each provenance element with its types where T is the set of element types as introduced by our provenance model,
- $attr : PE \rightarrow \mathfrak{P}(A \times V)$ is a mapping that associates each provenance element with additional properties represented by attribute-value pairs where A is a set of available attributes and V is a set of values.

We do not specify the sets A and V any further because the available attributes, possible values, and the meaning of these depend on the use case. However, we introduce an abbreviated notation to refer to the target of an edge in a provenance graph: if $(pe_1, pe_2, \mathrm{rn}) \in R$ we write $pe_1 \xrightarrow{\mathrm{rn}} \circ = pe_2$.

IV. Provenance-Based Quality Assessment

The method is based on provenance graphs represented using our provenance model. This approach should be regarded as a blueprint for the development of actual assessment methods that address specific scenarios and focus on specific quality criteria. This section introduces the general method and discusses questions that must be addressed when applying this method for a specific quality criterion.

A. The General Assessment Approach

The main idea of our approach is the automated determination of a quality measure for a data item, from so called *impact value*s, which represent the influence of the elements

in a provenance graph on the particular quality of the assessed data item. We divide the assessment procedure into three steps:

1) Generate a provenance graph for the data item;
2) Annotate the provenance graph with impact values;
3) Calculate an IQ-score for the data item from the annotated provenance graph.

In order to use the provenance of a data item for automated quality assessment this provenance has to be represented in the assessment system. We propose to use provenance graphs as introduced in Section III for this purpose. Hence, the first step of an assessment procedure must be the generation of such a graph for the data item that is to be assessed, i.e., the *considered data item*. This step comprises collecting the necessary provenance information about the data item.

Some, if not all, of the provenance elements might have had an influence on certain qualities of the assessed data item. Some of these influences are known to us; others are possible or cannot be ruled out. Both types of influences, known as well as possible influences, have an impact on our assessment of the qualities. We propose to represent this impact by *impact value*s associated with the corresponding provenance elements. For instance, the possibility of manipulating published data by a service provider may affect the believability and the assumed accuracy of the data; an impact value for a service provider could represent the provider's manipulation probability. An example for a known influence is the execution time of a data creation which has an impact on the timeliness assumed for the created data item. Notice, there can be different kinds of impact values for different types of provenance elements.

The second step of our assessment procedure comprises determining these impact values; the system adds annotations to the provenance graph generated from step 1, associating elements in the provenance graph with estimated impact values. Formally, an *annotated provenance graph* is a pair (pg, ann) where $pg = (PE, R, type, attr)$ is a provenance graph and $ann : PE \rightarrow \mathfrak{P}(I)$ is a mapping that associates a provenance element with a set of impact values; each impact value $(\mathsf{n}, v) \in I$ has a name n and the actual value v. For $(\mathsf{n}, v) \in ann(pe)$ we write $\mathsf{n}[pe] = v$.

In the final step the system executes a function to calculate a value that represents the information quality of the considered data item using the annotated provenance graph from step 2.

B. Designing Actual Assessment Methods

To apply our assessment approach one must first develop the presented method into an actual assessment method that is tailored for the quality criterion of interest. In the following we discuss design decisions that must be considered at each step and we specify the questions that must be addressed.

The most fundamental question that must be answered in the beginning is: *For which quality criterion do we want to apply the method?* This decision influences every aspect of an application of our approach. In the remainder we consecutively focus on the three steps of our assessment method. However, the design decisions for the three steps partly depend on each other. For this reason, designing an actual assessment method should be an iterative process.

Considering step 1 of the assessment method it is necessary to ensure the generation of a provenance graph that is suitable for the assessment. To specify suitability in the given context one has to ask: *What types of provenance elements are necessary to determine the considered information quality* and *what level of detail (i.e. granularity) is necessary to describe the provenance elements in the application scenario?* To answer these questions we propose to study the literature that deals with the considered quality criterion. A good starting point is Pipino et al. [13]. Based on the answers to the above two questions, the procedure for generating provenance graphs can be developed. Defining this procedure requires to address the question: *Where and how do we get the provenance information to generate the provenance graph for a data item?* Basically, there are two complementary options to obtain provenance information: some pieces of provenance information can be recorded by the system; for other pieces the system relies on meta-data provided by third parties. In [3] we discuss these options. Furthermore, we are working on the Provenance Vocabulary[1] to enable the publication of provenance-related metadata in the Web of data.

The fundamental questions that have to be answered for the implementation of step 2 are: *How might each type of provenance element influence the quality of interest* and *what kind of impact values are necessary for the application scenario?* The answers to these questions substantially depend on the considered quality criterion as well as on the assessment function used in step 3. Notice, impact values need not necessarily be numerical; they could also be of a more abstract nature such as the simple weighting "high impact". After specifying the impact values it is necessary to address the question: *How do we determine the impact values* or *where do we get them from?* Some of the impact values might already be part of the provenance information such as the creation time in the aforementioned timeliness example; others might be calculated based on the provenance graph. Certain kinds of impact values could also be determined based on user input. Another possibility is to estimate impact values by taking background knowledge about information consumers or providers into account. For instance, a data creator's credibility which influences believability assessments could be determined based on former experiences as well as on recommendations from other users.

The main questions regarding step 3 of the assessment method are: *How can we represent the considered information quality by a value* and *what function do we use to calculate such a value from the annotated provenance graph?* Again, answering these questions fundamentally depends on the quality criterion. The calculated value could be a single number in a specific interval; but it could also be a vector of numbers or an element of a set of discrete values. In any case, it is important to specify what such a value means. The

[1] http://purl.org/net/provenance/

definition of the applied function depends on the impact values introduced at step 2. For this reason, we recommend to develop the function together with specification of the impact values. For the development of this function it is important to bear in mind that the results of steps 1 and 2 cannot be guaranteed to be complete in many cases; the provenance graph could be fragmentary or some annotations could be missing due to the lack of certain information required during steps 1 and 2. Hence, the function for step 3 must not assume to operate on an ideal annotated provenance graph but it must be able to deal with incompleteness.

V. Provenance-Based Assessment of Timeliness

In this section we exemplarily apply our general assessment approach to assess the timeliness of data from the Web. We first give a brief introduction to timeliness and how it can be calculated; we then illustrate the design and the execution of the three steps of assessment; finally, we propose a way to deal with incomplete provenance information.

A. Representing and Calculating Timeliness

Timeliness is an intrinsic IQ criterion [14] that is often referred to as a task-dependend up-to-dateness of a data item [13], [15]. Ballou et al. represent timeliness by an absolute measure on a continuous scale from 0 to 1 where data with 1 "meet the most strict timeliness standard" [8] and 0 is unacceptable. This timeliness measure can be calculated using the following formula [8]:

$$Timeliness = (max(1 - Currency/Volatility, 0))^s \quad (1)$$

In this formula, *Volatility* is "the length of time the data remains valid" which is analogous to the shelf life of perishable products [8]; *Currency* is "the age of the data when it is delivered to the user" [15] which can be calculated according to [8] by the following formula:

$$Currency = Delivery\ Time - Input\ Time + Age \quad (2)$$

where *Delivery Time* is the time when the data was delivered to the user; *Input Time* is the time when the data was entered in the system; and *Age* is how old the data was at *Input Time*.

The exponent s in (1) is a parameter that controls the sensitivity of *Timeliness* to the *Currency-Volatility* ratio. The ratio should be large (e.g., $s = 2$) for highly volatile data and be small (e.g., $s = 0.5$) for long shelf life data [8].

Note that in Ballou et al.'s paper [8] the timeliness formula is defined in a closed "information manufacturing system", which processes primitive data units from outside. Hence, the semantics of *Age*, *Input Time*, and *Delivery Time* might be different w.r.t. to an open-world system, like the Web.

On the Web, we do not have primitive data from the *outside*. Instead, we have *unprocessed data* and *derived data*. Unprocessed data are data items for which the creation did not depend on other data items; i.e., no source data was used for their creation. Derived data, in contrast, was derived from other data items.

1) Timeliness of Unprocessed Data: For an unprocessed data item, its *Age* is 0 because it did not exist before; its *Input Time* is the time when its creation was finished; and its *Delivery Time* should be "now", i.e., the time when the timeliness of the *considered data item* is assessed. This means that the *Currency* values for unprocessed data items differ only by their creation time. To calculate the *Timeliness* of unprocessed data items using formula (1) we also need the *Volatility*. We could speak of volatility exclusively as *shelf life*, as Ballou et al. [8] do. Alternatively, we could speak of *expiry time* and adapt the formula from the Sampaio et al. [15]:

$$Volatility = Expiry\ Time - Input\ Time + Age \quad (3)$$

2) Timeliness of Derived Data: Ballou et al. compute the timeliness of data outputs from a *processing block* as a weighted average value [8]. In our method, for a derived data item, if it is caused by only one source data item, then it has the same timeliness value as the source data item; if it is caused by multiple source data items, then its timeliness value should be a weighted average of the timeliness values of the source data items.

B. Constructing the Provenance Graph

We adopt the calculation approach outlined in the previous section to apply our provenance-based assessment method for the determination of timeliness. The first step is to generate a provenance graph for the considered data item. For this work we assume the availability of all provenance information.

Example 1: We demonstrate the method applied to assessing timeliness of temperature measurements taken by a sensor. These measurements are unprocessed data items. They are taken every 1 hour, and they are stored in a Web-accessible storage device immediately. A system accesses these measurement from the storage device for further processing; in order to process the measures the system evaluates their timeliness.

We represent the provenance of a specific measure by a provenance graph $pe = (PE, R, type, attr)$ which is illustrated in Figure 1. PE contains the measure msr, the sensor $sens$, the data creation $cExc$ that produced msr, the storage device $stor$, the system sys, the data access $aExc$, and the document doc with which msr was retrieved during $aExc$. Given msr was taken at 10:00 and doc was retrieved at 10:13 it holds $attr(cExc) = \{(execTime, 10{:}00)\}$ and $attr(aExc) = \{(execTime, 10{:}13)\}$. □

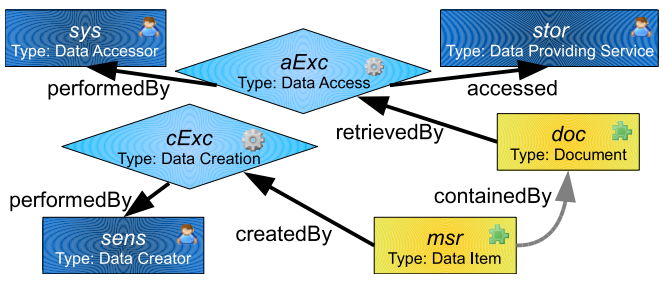

Fig. 1. Provenance graph representing our running example (cf. Example 1).

C. Adding Impact Values

The second step of the assessment method includes the annotation of the provenance graph with impact values. In order to design this step we study the relevance of different pieces of provenance information for the timeliness assessment. In particular, we discuss the relation of the provenance element types introduced by our provenance model to the calculation approach outlined before (cf. Section V-A).

Data creation executions have a direct influence on the timeliness assessment. As discussed before, the creation time of unprocessed data items corresponds to the input time in formula (2). Hence, we annotate each data creation element that is not associated with source data with a *creation time impact value*. It is not necessary to explicitly determine these kind of impact values because they are already represented in the provenance graph as an attribute of the data creation elements.

Data creations that yield a derived data item have an influence on the timeliness of this item if multiple source data items were used (cf. Section V-A2). We reflect this influence by another impact value: each data creation element that takes multiple source data items as inputs is annotated with a *weights impact value*. This impact value represents the weights that can be used to calculate the weighted average of the timeliness values of the source data items. Ballou et al. write: "The weights could reflect the size of the data units that are merged, their importance or some combination of attributes." [8]. In this paper, we leave the choice of the weights to applications adopting our assessment method, because this choice should be based on actual needs from their information consumers.

Creation guidelines may have an impact on IQ criteria such as accuracy and reliability. However, creation guidelines have no influence on the timeliness of the created data.

Source data may have an impact on the timeliness assessment. According to the calculation approach the timeliness of each derived data item can be ascribed to a combination of the timeliness values of unprocessed data items. Hence, in the ideal case of a complete provenance graph only the unprocessed data items have a direct influence on the timeliness of the considered data item. While their currency can be determined with the aforementioned creation time impact values we also need the volatility to calculate their timeliness using formula (1). To enable the calculation of their volatility using formula (3) we annotate each unprocessed data item with an *expiry time impact value*. We assume these impact values can be determined based on the input from users who configure a default expiry time for data from specific data creators or for data with a specific content.

Data creators have an influence on the volatility of unprocessed data items as discussed before. The previously mentioned strategy for determining the expiry time impact values reflects this influence.

Example 2: We annotate the provenance graph pg from Example 1 with impact values as follows. The data creation $cExc$ is not associated with source data; hence, it has to be annotated with a creation time impact value that refers to its $execTime$ attribute: $ann(cExc) = \{(\mathsf{crtT}, 10{:}00)\}$. Furthermore, pg contains the unprocessed data item msr which has to be annotated with an expiry time impact value. It is possible to determine this value based on the information that $sens$ takes the measures every hour. Hence, it holds: $ann(msr) = \{(\mathsf{expT}, 11{:}00)\}$. The other elements in pg do not have an influence on our timeliness assessment. □

In the ideal case of a complete provenance graph the elements that belong to the data access dimension can be ignored for the timeliness assessment. However, in the likely case that information about the creation of a (source) data item is missing or that it is impossible to determine one of the impact values introduced so far. Hence, we propose to consider the data access related elements as fall-back. For these data items, the *Input Time* in formulas (2) and (3) is the access time associated with the corresponding data access execution. Furthermore, the *Age* for these items is probably larger than 0, assuming that they, or the data from which they were derived, were not created at the time of the access. We annotate each of these data items with a *timeliness impact value* that represents a timeliness value estimated for them. This value could be estimated based on different data access related provenance elements. For instance, knowing when a data publisher updates her data may, in combination with the access time, be an indicator for the *Age*. The *Expiry Time* might be estimated based on information about the update frequency of the data publisher. After all, it must be realized that the timeliness impact values can only be estimates at best.

D. Calculating Timeliness

Based on the impact values in the annotated provenance graph it is possible to calculate timeliness by adopting formulas (1) to (3). The recursive function t in Figure 2 implements this idea: t incorporates (1) to (3) to calculate timeliness at step 3 of our assessment method. For a data item with incomplete provenance information t returns the timeliness impact value that is annotated to this item (cf. first case in the equation in Figure 2). For unprocessed data items t applies the formulas (1) to (3) using the corresponding creation time and expiry time impact values (cf. second case in the equation). For derived data items that were created with a single source data item t returns the timeliness value that is recursively calculated for the source data item (cf. third case). Finally, for other derived data items t uses the weights impact value of the corresponding data creation element to calculate a weighted average of the recursively calculated timeliness values of the source items (cf. fourth case).

Example 3: Based on the annotated provenance graph (pg, ann) from Example 2 it is possible to calculate the timeliness of msr. Since $msr \xrightarrow{\text{createdBy}} \circ \xrightarrow{\text{used}} \circ = \varnothing$ it holds:

$$t(msr) = \left(\max\Big(1 - \frac{now - \mathsf{crtT}\big[cExc\big]}{\mathsf{expT}\big[msr\big] - \mathsf{crtT}\big[cExc\big]}, 0\Big)\right)^s$$
$$= \left(\max\Big(1 - \frac{now - 10{:}00}{11{:}00 - 10{:}00}, 0\Big)\right)^s$$

$$t(d) = \begin{cases} \mathsf{timeliness}\big[d\big] & \text{if } d \overset{\text{createdBy}}{\longrightarrow} \circ \text{ is unknown,} \\ \left(\max\Big(1 - \frac{now - \mathsf{crtT}\Big[d \overset{\text{createdBy}}{\longrightarrow} \circ\Big]}{\mathsf{expT}\big[d\big] - \mathsf{crtT}\Big[d \overset{\text{createdBy}}{\longrightarrow} \circ\Big]}, 0\Big)\right)^{s} & \text{if } d \overset{\text{createdBy}}{\longrightarrow} \circ \overset{\text{used}}{\longrightarrow} \circ = \varnothing, \\ t(d_s) & \text{if } d \overset{\text{createdBy}}{\longrightarrow} \circ \overset{\text{used}}{\longrightarrow} \circ = \{d_s\}, \\ \frac{\sum_{d_s \in d \overset{\text{createdBy}}{\longrightarrow} \circ \overset{\text{used}}{\longrightarrow} \circ} \mathsf{weight}\Big[d \overset{\text{createdBy}}{\longrightarrow} \circ\Big]_{d_s} \cdot t(d_s)}{\sum_{d_s \in d \overset{\text{createdBy}}{\longrightarrow} \circ \overset{\text{used}}{\longrightarrow} \circ} \mathsf{weight}\Big[d \overset{\text{createdBy}}{\longrightarrow} \circ\Big]_{d_s}} & \text{if } \Big|d \overset{\text{createdBy}}{\longrightarrow} \circ \overset{\text{used}}{\longrightarrow} \circ\Big| > 1. \end{cases}$$

Fig. 2. The recursive function that calculates the timeliness of a data item d based on impact values from the annotated provenance graph for d.

Given $s = 1$ and the timeliness assessment happens at 10:15, i.e. $now =$ 10:15, we get the result:

$$= \max\Big(1 - \frac{0.25\text{h}}{1\text{h}}, 0\Big) = \underline{0.75} \qquad \square$$

E. Dealing with Incomplete Provenance Information

Our timeliness assessment method deals with incomplete information by using alternative impact values; furthermore, certain impact values can only be determined by estimation. Thus, the calculated timeliness value becomes an approximation rather than an exact assessment. To make the degree of inexactness explicit we propose to associate the calculated timeliness value with a certainty value. This additional value represents the certainty of whether the calculated timeliness is correct. We suggest to represent certainty with a value in the interval [0,1] where 1 represents absolute certainty, i.e. no doubt, and 0 represents absolute unvertainty, i.e. the calculated timeliness value is useless. In the following we outline an approach to calculate the certainty value during the execution of our assessment method.

We assume a value, initialized to 1, that is accessible throughout the whole assessment procedure. During the execution of steps 1 and 2 this value is incrementally decreased whenever i) a part of the provenance graph cannot be generated appropriately due to missing provenance information and whenever ii) an impact value is estimated. With each decrease the value should be reduced by a certain percentage where the extent of this percentage may differ for different decreases. Identifying appropriate extents is subject to further research. For instance, in the case of missing provenance information the importance of this information to the assessment may affect the amount of reduction. Decreases due to impact value estimation may differ depending on the reliability of the applied estimation strategy. However, after the completion of step 2 the decreased value represents the reliability of the annotated provenance graph. Since the calculation in step 3 is solely based on this graph the value also represents a certainty regarding the correctness of the calculated timeliness value.

VI. Conclusion

In this paper we propose a provenance model for Web data provenance and an assessment method for evaluating the quality of data on the Web using provenance graphs based on this model. Our provenance model introduces a new dimension of provenance information, i.e. the provenance of data access, to the existing provenance research. We are gathering feedback to our model from different communities and we foresee continuing development of our provenance vocabulary driven by well-defined use cases. In this paper, we demonstrate assessing the timeliness of data on the Web using our method. We plan to implement this method as part of a Web data publication framework in the near future and to apply this method to the assessment of other quality criteria, such as accuracy. Our method should be generic enough to incorporate subjective quality indicators derived from Web data provenance. Existing work on evaluating and filtering subjective quality indicators will be considered and appropriately applied.

VII. Acknowledgement

Part of this work is funded by EPSRC Postdoctoral Fellowship to Dr Jun Zhao (Grant EP/G049327/1).

References

[1] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *Int. Journal on Semantic Web and Information Systems, Special Issue on Linked Data*, 2009, in press.

[2] W. L. Yang, M. S. Diane, B. K. Kahn, and Y. W. Richard, "AIMQ: a methodology for information quality assessment," *Information & Management*, vol. 40, no. 2, 2002.

[3] O. Hartig, "Provenance Information in the Web of Data," in *Proc. of the Linked Data on the Web Workshop at WWW*, 2009.

[4] F. Naumann, *Quality-driven query answering for integrated information systems*. Springer Verlag, 2002.

[5] M. Bobrowski, M. Marré, and D. Yankelevich., "A homogeneous framework to measure data quality," in *Proc. of IQ*, 1999.

[6] A. Motro and I. Rakov, "Estimating the quality of databases," in *Proc. of FQAS*, 1998.

[7] J.-R. Gruser, L. Raschid, V. Zadorozhny, and T. Zhan, "Learning response time for websources using query feedback and application in query optimization," *VLDB Journal*, vol. 9, no. 1, 2000.

[8] D. Ballou, R. Wang, H. Pazer, and G. K. Tayi, "Modeling Information Manufacturing Systems to Determine Information Product Quality," *Management Science*, vol. 44, no. 4, 1998.

[9] S. C. Wong, S. Miles, W. Fang, P. Groth, and L. Moreau, "Provenance-based validation of e-science experiments," in *Proc. of ISWC*, 2005.

[10] J. Golbeck and A. Mannes, "Using Trust and Provenance for Content Filtering on the Semantic Web," in *Proc. of the Models of Trust for the Web Workshop at WWW*, 2006.

[11] Y. Simmhan, B. Plale, and D. Gannon, "A Survey of Data Provenance in e-Science," *SIGMOD Record*, vol. 34, no. 3, 2005.

[12] W. C. Tan, "Provenance in Databases: Past, Current, and Future," *IEEE Data Engineering Bulletin*, vol. 30, no. 4, 2007.

[13] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data Quality Assessment," *Communications of the ACM*, vol. 45, no. 4, 2002.

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

[15] S. de F. Mendes Sampaio, C. Dong, and P. Sampaio, "Incorporating the Timeliness Quality Dimension in Internet Query Systems," in *Proc. of WISE*, 2005.