Data Provenance and Believability in the Semantic Web

Nicolas Prat
Stuart Madnick

Working Paper CISL# 20010-06

June 2010

Composite Information Systems Laboratory (CISL)
Sloan School of Management, Room E53-320
Massachusetts Institute of Technology
Cambridge, MA 02142

Data Provenance and Believability in the Semantic Web

Nicolas Prat
ESSEC Business School
Avenue Bernard Hirsch
BP 50105 – 95021 Cergy Cedex - FRANCE
prat@essec.fr

Stuart Madnick
MIT Sloan School of Management
30 Wadsworth Street – Room E53-321
Cambridge MA 02142 - USA
smadnick@mit.edu

ABSTRACT

In the semantic Web, human and software agents can link data and combine them to create new data. In this context, the representation and management of data provenance is crucial. Data provenance helps determine the believability of data, which is an important aspect of data quality. This paper focuses on the representation of data provenance in the Web of linked data, and the use of provenance information to measure the believability of data. We present our provenance model and investigate how the concepts of this model can be represented in the semantic Web using existing languages and vocabularies to measure data believability. To our knowledge, our approach, applied in this paper, is the first to develop a precise, systematic approach to measuring data believability and making explicit use of provenance-based measurements in the Web of linked data.

1. INTRODUCTION

In the semantic Web, human and software agents can link data and combine them to create new data. The term "linked data" has been introduced to refer to a set of best practices for publishing and connecting structured data on the Web; the Linking Open Data project (http://linkeddata.org/) aims at bootstrapping the semantic Web by publishing datasets on the Web and linking them with other datasets.

The representation and management of data provenance is crucial [8,16] because it helps determine the trustworthiness and, more generally, the believability of data. Data believability is an important aspect of data quality and has been defined as "the extent to which data are *regarded* as true, real and credible" [19]. Believability can be considered as synonymous to credibility [18]. We argue that it is crucial to automate provenance-based measurement of data believability as much as possible, not only for reasons of scalability, but also to reduce the risks of "incredulity errors" and "gullibility errors" [18]. Incredulity errors happen when the product (in our case, data) is credible but the user perceives it as not credible. Conversely, gullibility errors happen when the product is not credible but the user perceives it as credible.

In this paper, we focus on the representation of provenance and the automated, provenance-based computation of data believability. In contrast, in previous work [13,14], we have proposed a provenance model, as well as metrics and a computation approach to evaluate data believability based on provenance information. In this paper, we refine our approach and operationalize it in the specific context of linked data. Beyond the specific application to linked data, our work is at the intersection of provenance research and data quality research.

The issue of provenance has been investigated in several domains. Provenance (aka lineage) is defined in [17] as "information that helps determine the derivation history of a data product, starting from its original sources". Provenance is a key research issue in database research [1] and e-science [17], as well as other domains. Among the several applications of data provenance, provenance information is crucial for users to decide to what extent they can *believe* the electronic data [11]. However, the literature on data provenance lacks a global, precise approach to computing data believability based in provenance information.

In data quality research, believability (aka credibility) appears as a dimension of data quality [3]. The survey by R. Wang and D. Strong [19] shows that data consumers consider believability as an important aspect of data quality. However, the data quality literature lacks metrics for precisely computing data believability, which consists in several sub-dimensions. Guidelines for measuring data believability may be found in [10] (pp. 57-58). However, these guidelines remain quite general and no formal metrics are proposed. An earlier data quality paper [2] addresses the issue of lineage-based data quality assessment (even if the terms of lineage or provenance are not explicitly mentioned). However, the authors address data quality (defined as the absence of errors) in a general and syntactic way. We argue that the different dimensions of quality (and, more particularly, of believability) have different semantics, which should be explicitly considered for quality computation.

To our knowledge, this work, based on our earlier work reported in [13,14], is the first one to develop a precise, systematic approach to measuring data believability and making explicit use of provenance-based measurements. In this paper, we refine our earlier approach and apply it to the context of the Web of linked data.

The paper is structured as follows. Section 2 presents the sub-dimensions of data believability. Section 3 presents the provenance model. This model aims at representing and capturing the information that will subsequently be used for automatic computation of data believability. The model is independent of the implementation context (e.g., relational database, semantic Web, etc.) Section 4 investigates how this provenance model can be operationalized in the context of the Web of linked data. To this end, we map the concepts of the provenance model with concepts of RDF vocabularies and languages. Section 5 presents our approach for provenance-based believability computation, applying the approach to a concrete example. Section 6 concludes and points to further research.

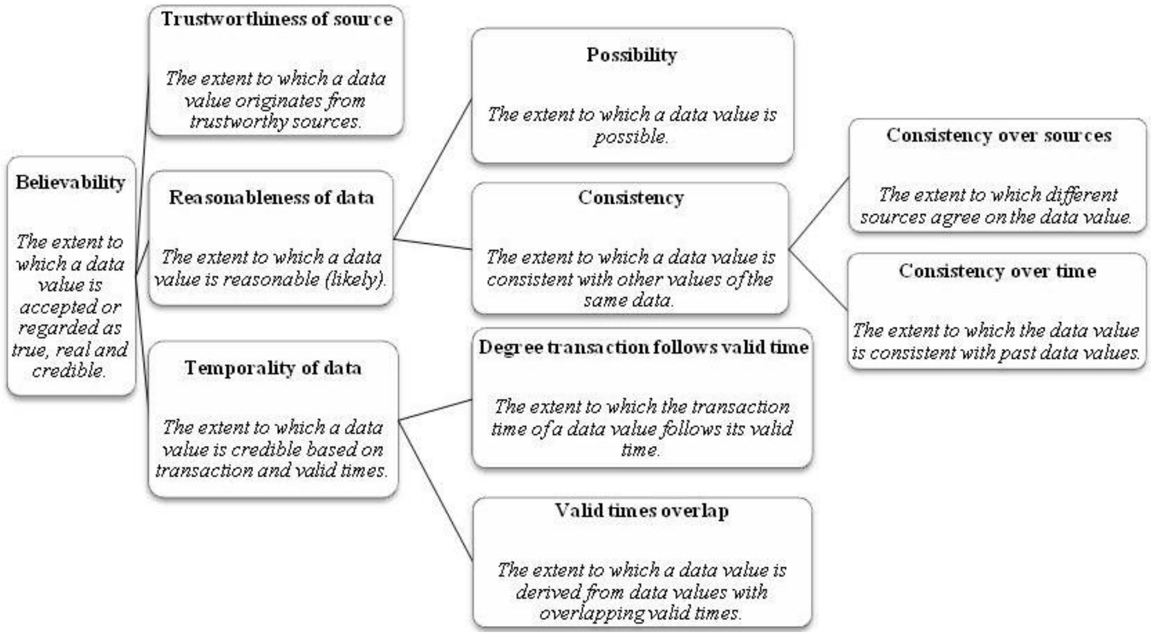
Believability
The extent to which a data value is accepted or regarded as true, real and credible.
Trustworthiness of source
The extent to which a data value originates from trustworthy sources.
Reasonableness of data
The extent to which a data value is reasonable (likely).
Temporality of data
The extent to which a data value is credible based on transaction and valid times.
Possibility
The extent to which a data value is possible.
Consistency
The extent to which a data value is consistent with other values of the same data.
Degree transaction follows valid time
The extent to which the transaction time of a data value follows its valid time.
Valid times overlap
The extent to which a data value is derived from data values with overlapping valid times.
Consistency over sources
The extent to which different sources agree on the data value.
Consistency over time
The extent to which the data value is consistent with past data values.

Figure 1. Sub-dimensions of believability

2. BELIEVABILITY SUB-DIMENSIONS

A dimension of data quality, believability is itself a multidimensional concept. [10] proposes three sub-dimensions of believability: (1) of source, (2) compared to internal commonsense standard, and (3) based on temporality of data. In our approach, we refine this typology [13], decomposing the three initial sub-dimensions of believability. Figure 1 illustrates our ontology of the sub-dimensions of data believability. Although believability is often related to trust [18], trustworthiness is only one of the several aspects of believability.

Our provenance model is aimed at representing and capturing believability-related provenance information, i.e. the information that will be used to compute the different sub-dimensions of data believability. This model is presented in the next section.

3. PROVENANCE MODEL

Several provenance models have been proposed in the literature [4,11,15,20]. The model proposed in [8] is dedicated to provenance capture in the context of linked data. Compared with these models, our provenance model was designed with the specific objective of computing the different sub-dimensions of believability. Figure 2 shows our model, represented as a class diagram in UML notation [12]. An earlier version of the model is presented in [14].

Since our goal is to assess the believability of data values, they are the central concept of the model. A data value may be atomic or complex (e.g. relational records, XML files, etc.) Our current research focuses on atomic, numeric data values.

Source vs. Resulting data value. A data value (e.g. 109 900 000 000) is the instance of a data (e.g. “the public expenditure in health in the UK, in financial year 2008-2009, in £”). A data value may be a source, or a resulting data value (output of a process run). We introduce this distinction between source and resulting data values because we use different believability metrics for these two types of values. The notion of source data value is relative to the information system under consideration: very often, a “source” data value is itself the result of process runs, but these processes are outside the scope of the information system.

A process run is the instantiation (i.e. execution) of a process. This distinction between process runs and processes parallels the distinction between data values and data, respectively. Processes may have several inputs but only have one output.

VLDB '10, September 13-17, 2010, Singapore.

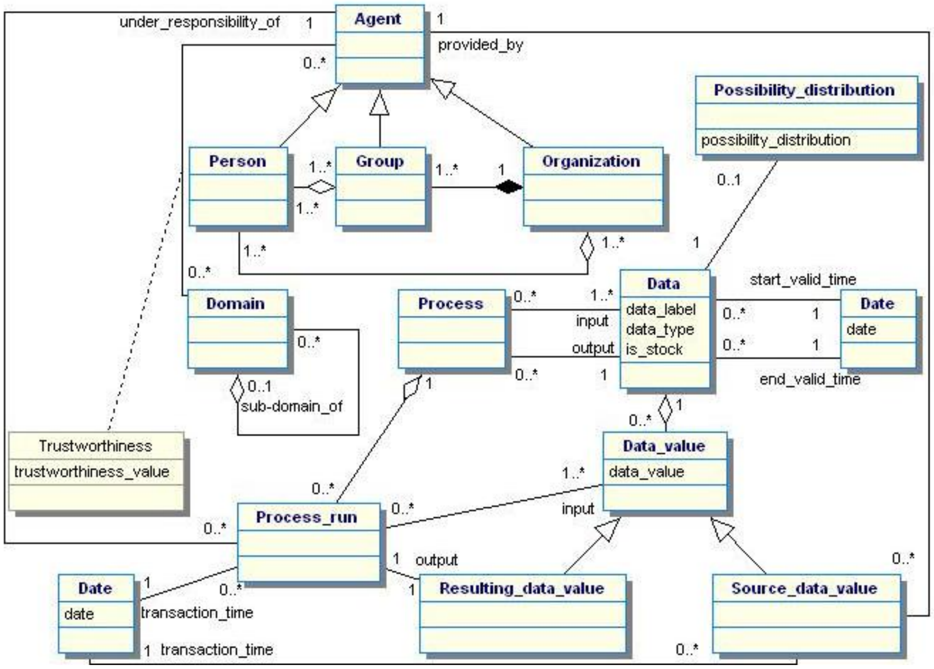

Figure 2. Provenance model

Transaction time vs. Valid time. Every data value has a transaction time. For a resulting data value, the transaction time is the execution time of the process run that generated the data value. For a source data value, the transaction time is attached directly to the data value. In addition to transaction time, we use the notion of valid time, defined as follows in [9] (p. 53): “The valid time of a fact is the time when the fact is true in the modeled reality. A fact may have associated any number of instants and time intervals, with single instants and intervals being important special cases.” In other words, valid time is the period of interest of data, while transaction time is when the data value was computed. Contrary to transaction time which depends on process execution, valid time depends on the semantics of data. For example, for the data “the public expenditure in health in the UK, in financial year 2008-2009, in British pounds”, the start valid time is April 1, 2008 and the end valid time is March 31, 2009. The distinction between valid time and transaction time is used explicitly in the assessment of the two sub-dimensions of temporality. Although transaction time and valid time are standard concepts in temporal databases, we haven’t encountered this distinction in extant provenance models.

Related to the concept of valid time is the distinction between stocks and flows. This distinction is made in several disciplines, including system dynamics [7]. In accounting, an account balance at a given time is an example of stock, while sales over a period of time is a flow. The valid time of a flow is an interval of time, whereas the valid time of a stock is normally an instant (although, in some cases, the value of a stock may be represented as an average over a period of time). Consequently, our metrics for computing believability deal with stocks and flows differently.

Possibility. When computing data believability, we will use the concept of possibility defined in possibility theory [5]. Accordingly, a possibility distribution is associated with data. Possibility distributions take their values between 0 (impossible) and 1 (totally possible) and may be defined on intervals. For example, if one considers that “the public expenditure in health in the UK, in financial year 2008-2009, in British pounds” is somewhere between 50 000 000 000 and 150 000 000 000, this can be expressed by a possibility distribution with a value of 1 in the [50 000 000 000 , 150 000 000 000] interval, and 0 outside. In this case, the possibility distribution is equivalent to an integrity constraint stating that the value should be in the [50 000 000 000 , 150 000 000 000] range. However, possibility distributions allow for a fine-tuned representation of uncertainty, by using possibility values between 0 and 1.

Trustworthiness. Processes are under the responsibility of agents (organizations, groups or persons). This concept also represents the providers of the source data values. When computing believability, we are not interested in agents per se, but in the trustworthiness of these agents. The concept of trustworthiness is essential for assessing the dimension “trustworthiness of source”. Trustworthiness is evaluated for an agent, for a specific knowledge domain. Trustworthiness in an agent for a domain is measured by a trustworthiness value, normalized between 0 and 1. We assume that these values are obtained from outside sources, e.g. review systems. For example, the trustworthiness of the organization “the Economist Magazine” in the domain of economics is available from Epinions (www.epinions.com).

In the appendix, a table illustrates the relationship between our provenance model and the computation of believability, by showing which concepts are used in which metrics. The concepts of the provenance model and their properties are shown in the left column (properties are properties of classes or roles of relationships; when roles have not been named, we use the name of the target class, e.g. Person.groups represents the groups to which a person belongs). The believability metrics are shown in the right column (they correspond to the terminal nodes of the ontology of believability sub-dimensions in Figure 1).

In the next section, we investigate how the concepts of our provenance model can be operationalized in the Web of linked data, by mapping the provenance model with RDF vocabularies and languages.

4. MAPPING THE PROVENANCE MODEL INTO RDF

RDF is central to the Web of linked data. It is used to represent datasets as resources and properties, and connect datasets together.

Several RDF vocabularies are available. In this paper, our purpose is not to propose yet another vocabulary, but rather to investigate how extant vocabularies and languages can be used to operationalize our provenance model. To this end, we map the concepts and properties of the provenance model into RDF. The RDF vocabularies and languages that we have found relevant for this purpose are: XML Schema (http://www.w3.org/TR/xmlschema-2/), OWL (http://www.w3.org/TR/owl-overview/), OWL-S (http://www.w3.org/Submission/OWL-S/), Dublin Core (http://dublincore.org/documents/dcmi-terms/), FOAF (Friend Of A Friend, http://xmlns.com/foaf/spec/20100101.html), SKOS (http://www.w3.org/TR/2009/REC-skos-reference-20090818/), the Trust Ontology (http://trust.mindswap.org/trustOnt.shtml), SCOVO (Statistical Core Vocabulary, http://purl.org/NET/scovo#) and Data-gov (http://data-gov.tw.rpi.edu/).

The following tables show the mappings between the concepts (classes) and properties of our provenance model and classes and properties in these RDF vocabularies and languages. We divide our provenance model into three clusters: (1) agents, domains and trustworthiness, (2) data and (3) data values and processes.

4.1 Agents, Domains and Trustworthiness

Table 1. Mapping the provenance model into RDF: agents, domains and trustworthiness

Concept of the provenance model	RDF term	RDF vocabulary
Agent	Agent	FOAF
Person	Person	FOAF
• groups	member	FOAF
• organizations		
Group	Group	FOAF
• organization		
Organization	Organization	FOAF
Domain	Concept	SKOS
• sub_domain_of	narrower	SKOS
Trustworthiness	TopicalTrust	Trust Ontology
• agent	trustedPerson	Trust Ontology
• domain	trustSubject	Trust Ontology
• trustworthiness value	trustValue	Trust Ontology

The concepts of agent, person, group and organization in the provenance model can be mapped directly to the FOAF vocabulary, although (unlike FOAF) we exclude software agents from our definition of agents: even if a process may be executed by software, the agent responsible for a process run or providing a data value is human.

Despite this mapping of concepts between our model and FOAF, FOAF only partly represents membership information between persons, groups and organizations.

To represent knowledge domains and their hierarchical organization, SKOS is appropriate. This vocabulary has been designed to represent thesauri, taxonomies, classification schemes and subject heading systems.

Finally, the Trust Ontology is appropriate for representing the concept of trustworthiness and its properties.

4.2 Data

Table 2. Mapping the provenance model into RDF: data

Concept of the provenance model	RDF term	RDF vocabulary
Data	RDF property	(general case)
	Item	SCOVO
• data_label	Name of the RDF property	(general case)
	Name of the Item	SCOVO
• data_type	datatype	XML Schema
• is_stock		
• start_valid_time	time_period	Data-gov
	min	SCOVO
• end_valid_time	time_period	Data-gov
	max	SCOVO
Possibility_distribution		
• possibility_distribution	minInclusive minExclusive maxInclusive maxExclusive	XML Schema

In most RDF vocabularies, data are not represented as first-class citizens, but as properties of RDF classes. A notable exception is SCOVO, where the class Item represents a single piece of data. This allows for a richer, more precise representation of the different dimensions and properties of data.

The XML schema vocabulary can be used for data types (in our case, we focus on atomic, numeric data types).

Data-gov has a time_period property, defined as the "date or time interval(s) for which the data set provides data". This property corresponds to the concept of valid time. However, it is defined at the granularity level of a data set, and not for a specific piece of data. In SCOVO, a valid time can be associated with a specific piece of data (Item in SCOVO vocabulary): an Item has several dimensions, one of which can be the valid time, with a min (start_valid_time) and a max (end_valid_time).

We did not find a direct equivalent of possibility distributions in RDF vocabularies. However, XML Schema enables the definition of a lower bound and an upper bound for ordered domains of values. These lower and upper bounds define the interval outside of which the possibility value is equal to 0.

4.3 Data Values and Processes

Table 3. Mapping the provenance model with RDF: data values and processes

Concept of the provenance model	RDF term	RDF vocabulary
Data_value	RDF property value	(general case)
	rdf:value	SCOVO
• data_value	Value of the RDF property	(general case)
	rdf:value	SCOVO
• data	RDF property	(general case)
	Item	SCOVO
Source_data_value		
• provided_by	publisher creator contributor	Dublin Core
	agency	Data-gov
• transaction_time	created modified	Dublin Core
	date_released date_updated	Data-gov
Resulting_data_value		
• process_run	collection_mode statistical_ methodology	Data-gov
Process_run		
• under_responsibility_of	publisher creator contributor	Dublin Core
	agency	Data-gov
• transaction_time	created modified	Dublin Core
	date_released date_updated	Data-gov
• process		
Process	Process	OWL-S
• input	hasInput	OWL-S
• output	hasOutput	OWL-S

As pointed out in [8], Dublin Core can be used to represent data providers and time information (time of data creation or modification, i.e. transaction time). Dublin Core even distinguishes between the roles of publisher, creator and contributor. However, with Dublin Core alone, these properties cannot be represented at the level of the data value. Similarly to Dublin Core, Data-gov has the concepts of agency ("the government agency publishing the data set"), date_released and date_updated. These properties are defined at the granularity level of data sets.

OWL-S enables the representation of processes, their inputs and outputs. However, at the instance level, i.e. to represent the process execution from which a data value results, the RDF vocabularies do not provide a direct, satisfying solution. Data-gov has the properties of collection_mode and statistical_methodology, but these properties are not formalized and may have very different values. Examples of collection modes include radar, satellite, numeric prediction models… The statistical methodology is defined as "a description of the overall approach used for statistical design, sampling, data collection, statistical analysis, and estimation." An example of statistical methodology is "1 percent random, representative sample of administrative records of Social Security beneficiaries".

Summing up the mapping between our provenance model and RDF vocabularies and languages, extant vocabularies and languages provide several classes and properties for representing the concepts of our provenance model. However, in some cases we have no equivalent, and we need to use several vocabularies in conjunction (as often happens in extant data sets).

5. PROVENANCE-BASED BELIEVABILITY COMPUTATION

We illustrate our approach for provenance-based believability computation, operationalized in the Web of linked data. We first introduce the scenario. We then present metrics for the sub-dimensions of believability. We have defined a metric for each elementary sub-dimension of believability (i.e. for each terminal node in Figure 1); due to space limitation, we focus here on the metrics defined for the temporality of data. Finally, we present and illustrate our approach for spatio-temporal assessment of data believability (i.e. overal assessment of believability, based on the sub-dimensions of believability and the lineage of data values).

5.1 Example Scenario

The example is based on linked data from www.data.gov.uk, a recent initiative of the UK Government, advised by Sir T. Berners-Lee and N. Shadbolt. The data used in the example are presented in the lower part of Figure 3, along with provenance information. We assume that there is a need to compute the public expenditure in health and social protection per inhabitant in the UK, in 2008-2009, in USD (the data value v). As shown in Figure 3, this data value is computed from three data values: the public expenditure (in British pounds) in health and social protection in the UK in 2008-2009 (v_{21}), the population of the UK (v_{22}) and the exchange rate of the British pound (v_{23}). Values v_{22} and v_{23} are taken from the CIA World Fact Book, available on the Web. Value v_{21} is the sum of values v_{11} and v_{12}, which are stored as linked data in Data.gov.uk. Figure 3 shows the information that we will need to compute the believability of the final data value (the value v). This information is represented in the provenance model, and extracted from the Web (apart from the transaction times of data values v_{21} and v). Since this paper focuses on metrics related to the temporality of data, we focus on the information that we will need to compute these metrics. It should be noted that v_{22} and v_{23} represent stocks, while other values are flows.

The RDF representation uses the SCOVO vocabulary, representing, among other things, the start and end valid times for the dimension "financial year 2008-09". Figure 3 also shows the cell value corresponding to v_{11}, as it appears in Excel.

5.2 Metrics for Assessing the Sub-Dimensions of Believability

As a first building block of our approach for believability computation, we have defined 6 metrics. These metrics are used to assess the believability of each data value, for each elementary sub-dimension of believability. The metrics have been presented in [14]. Some metrics need to distinguish between source data values (in our case, v_{11}, v_{12}, v_{22} and v_{23}) and data values resulting from a process run (v_{21} and v).

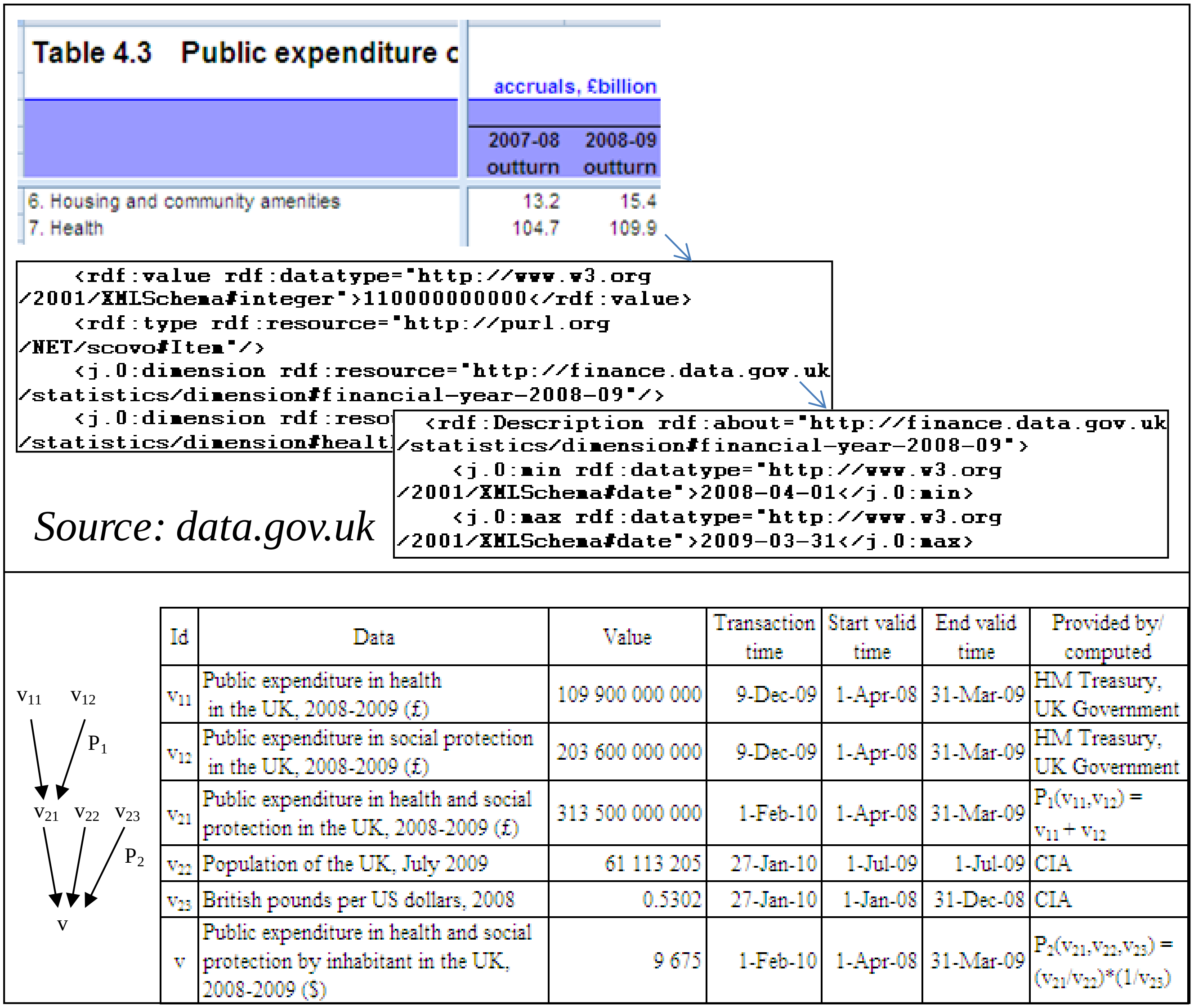

Table 4.3 Public expenditure c	accruals, £billion	
	2007-08 outturn	2008-09 outturn
6. Housing and community amenities	13.2	15.4
7. Health	104.7	109.9

```
    <rdf:value rdf:datatype="http://www.w3.org
/2001/XMLSchema#integer">110000000000</rdf:value>
    <rdf:type rdf:resource="http://purl.org
/NET/scovo#Item"/>
    <j.0:dimension rdf:resource="http://finance.data.gov.uk
/statistics/dimension#financial-year-2008-09"/>
    <j.0:dimension rdf:reso
/statistics/dimension#healt
```

```
  <rdf:Description rdf:about="http://finance.data.gov.uk
/statistics/dimension#financial-year-2008-09">
    <j.0:min rdf:datatype="http://www.w3.org
/2001/XMLSchema#date">2008-04-01</j.0:min>
    <j.0:max rdf:datatype="http://www.w3.org
/2001/XMLSchema#date">2009-03-31</j.0:max>
```

Source: data.gov.uk

Id	Data	Value	Transaction time	Start valid time	End valid time	Provided by/ computed
v_{11}	Public expenditure in health in the UK, 2008-2009 (£)	109 900 000 000	9-Dec-09	1-Apr-08	31-Mar-09	HM Treasury, UK Government
v_{12}	Public expenditure in social protection in the UK, 2008-2009 (£)	203 600 000 000	9-Dec-09	1-Apr-08	31-Mar-09	HM Treasury, UK Government
v_{21}	Public expenditure in health and social protection in the UK, 2008-2009 (£)	313 500 000 000	1-Feb-10	1-Apr-08	31-Mar-09	$P_1(v_{11},v_{12}) = v_{11} + v_{12}$
v_{22}	Population of the UK, July 2009	61 113 205	27-Jan-10	1-Jul-09	1-Jul-09	CIA
v_{23}	British pounds per US dollars, 2008	0.5302	27-Jan-10	1-Jan-08	31-Dec-08	CIA
v	Public expenditure in health and social protection by inhabitant in the UK, 2008-2009 ($)	9 675	1-Feb-10	1-Apr-08	31-Mar-09	$P_2(v_{21},v_{22},v_{23}) = (v_{21}/v_{22})*(1/v_{23})$

Figure 3. Example scenario

Our notation convention for metrics is based on the order believability sub-dimensions appear in Figure 1. For example, Q_{32} is the metric for valid times overlap. In this paper, we focus on the two metrics pertaining to the temporality of data (Q_{31} and Q_{32}). Compared with the initial version [14], these metrics have been significantly extended and clarified. We present the metrics and apply them to the example scenario.

Let us first consider Q_{31}(val), the metric for assessing the sub-dimension "**degree transaction follows valid time**" for a data value val. For this metric, the intuition is that a data value computed in advance (estimation) is more reliable as the valid time (especially the end valid time) of the data value approaches. When transaction time is equal or superior to the end valid time, it may also happen that the data value remains an estimation for a while (for example, in our scenario, values for recent financial years are accruals, as opposed to actual cash spending for earlier years). To capture this idea, we need a function that grows exponentially for transaction times before the end valid time, and also for transaction times after the end valid time. Thus, we need a function with a "S" shape (sigmoid function), bounded by 0 and 1. We can use the logistic function, defined as:

$$f(x) = \frac{1}{1+e^{-x}}$$

To control the shape of the curve, we can apply a decline coefficient, as proposed in the literature (e.g. in [6]). Since the exponential growth of the function is likely to be more rapid after the end valid time, we can define two logistic functions (each with a specific decline coefficient): one for transaction time before valid time, and one for transaction time after valid time.

Formally, the metric Q_{31}(val) is expressed as follows:

```
Let $tt$:Date such that $val.transaction\_time = tt$
Let $evt$:Date such that $val.data.end\_valid\_time = evt$
Let $\alpha1$ and $\alpha2$ be two decline factors ($\alpha2>\alpha1>0$)
        If ($tt<evt$)
        Then
```

```
$Q_{31}(val) = \frac{1}{1+e^{\alpha 1*(evt-tt)}}$
    Else
$Q_{31}(val) = \frac{1}{1+e^{\alpha 2*(evt-tt)}}$
    Endif
```

To illustrate the behavior of this metric, Figure 4 shows how the value of the metric evolves as a function of transaction time. We assume here that α1=0.005, α2=0.010, and the end valid time is March 31, 2009.

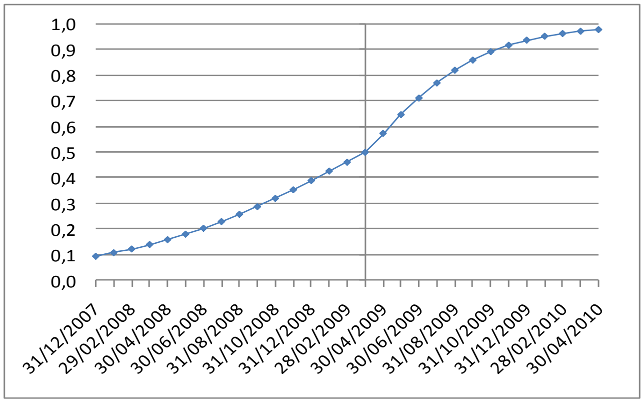

Figure 4. Degree transaction meets valid time: illustration of the metric

Let us now consider Q_{32}(val), the metric for assessing the sub-dimension "**valid times overlap**" for a data value val. This metric is defined only for resulting data values (in our example, v_{21} and v). For a data value val resulting from a process P with input values v_i, valid times overlap measures the extent to which the valid times of the v_i are consistent with each other, i.e. their degree of overlap. In case there is only one input value, valid times overlap is equal to 1. Otherwise, the idea of the metric is to compute, pair-wise, the valid times overlap of input data values, and compute the average. We distinguish between the cases when the valid time is an interval, and is an instant. By convention, we also represent instants as specific kinds of intervals, where the left bound is the same as the right bound, and the length of these intervals is one unit of time (e.g. the length of [1-Jul-09,1-Jul-09] is one day). When computing the overlap between an interval and an instant, we consider that the overlap is higher when the instant is located close to the middle of the interval (e.g. if we divide the yearly public expenditure in a country by the total population for that country at a given time, it is preferable to take the figure for the total population at the middle of the considered year; this figure is more likely to reflect the average population of the country for that year).

Formally, the metric Q_{32}(val) is expressed as follows:

```
Call n the number of input parameters of the process from which the value val results.
    If n=1
    Then $Q_{32}(val) = 1$
    Else
Call $v_i$ the $i^{th}$ input parameter of the process.
Call Min_valid_time (resp. Max_valid_time) the earliest (respectively latest) start valid time (respectively end valid time) among the valid times of the $v_i$.
Call $VTv_i$ the valid time interval of $v_i$ (interval [$svt_i$,$evt_i$], delimited by the start and end valid time of $v_i$; if $VTv_i$ is an instant, $svt_i$=$evt_i$).
$Q_{32}(val) = \frac{2}{n*(n-1)} *$
$(\sum_{i>j} \frac{overlap(i,j)}{length(\ Min\_valid\_time, Max\_valid\_time\ )})$
    Endif
overlap(i,j) is defined as follows:
    If ($VTv_i \cap VTv_j$ = ø)
    Then ovelap(i,j)=0
    Else
        If $VTv_i$ is not an instant and $VTv_j$ is an instant
        Then
$overlap(i,j) = \frac{\min(length(\ svt_i, evt_j\ ), length(\ svt_j, evt_i\ ))}{0.5*length(VTv_i)}$
        Endif
        If $VTv_i$ is an instant and $VTv_j$ is not an instant
        Then
$overlap(i,j) = \frac{\min(length(\ svt_j, evt_i\ ), length(\ svt_i, evt_j\ ))}{0.5*length(VTv_j)}$
        Endif
        If ($VTv_i$ and $VTv_j$ are not instants) or ($VTv_i$ and $VTv_j$ are instants)
        Then
$overlap(i,j) = length(VTv_i \cap VTv_j)$
        Endif
    Endif
```

Table 3 applies the two temporality metrics (Q_{31} and Q_{32}) to our example scenario.

Table 4. Temporality of data: application of the metrics

Id	Q_{31}	Q_{32}
v_{11}	0.926	
v_{12}	0.926	
v_{21}	0.956	1.000
v_{22}	0.891	
v_{23}	0.981	
v	0.956	0.167

$$\begin{array}{c|cccccc} & v_{11} & v_{12} & v_{21} & v_{22} & v_{23} & v \\ \hline v_{11} & 0 & 0 & a=0.35 & 0 & 0 & 0 \\ v_{12} & 0 & 0 & b=0.65 & 0 & 0 & 0 \\ v_{21} & 0 & 0 & 0 & 0 & 0 & c=0.33 \\ v_{22} & 0 & 0 & 0 & 0 & 0 & d=0.33 \\ v_{23} & 0 & 0 & 0 & 0 & 0 & e=0.33 \\ v & 0 & 0 & 0 & 0 & 0 & 0 \end{array}$$

$$\begin{array}{cccccc} v_{11} & v_{12} & v_{21} & v_{22} & v_{23} & v \\ \frac{\gamma^2 * a * c}{\sum} & \frac{\gamma^2 * b * c}{\sum} & \frac{\gamma * c}{\sum} & \frac{\gamma * d}{\sum} & \frac{\gamma * e}{\sum} & \frac{1}{\sum} \\ \uparrow & \uparrow & \uparrow & \uparrow & \uparrow & \uparrow \\ 0.02 & 0.03 & 0.11 & 0.11 & 0.11 & 0.63 \end{array}$$

$$\underline{NOTE}: \sum = \gamma^2 * a * c + \gamma^2 * b * c + \gamma * c + \gamma * d + \gamma * e + 1$$

Figure 5. Application scenario: matrix M(v) (left) and influence vector G(v) (right)

In this case, for the metric "degree transaction follows valid time" (Q_{31}), we have taken the decline factors $\alpha 1=0.005$, and $\alpha 2=0.010$, for all data values. We could choose different decline factors for the different data, reflecting the temporal behavior of data (for example, the evolution of population and of currencies each have their specific behavior).

In Table 3, the metric for valid times overlap (Q_{32}) is defined for v_{21} and v only. The metric reflects perfect overlap for v_{21} (input values v_{11} and v_{12} have the same valid times). In the case of v, input values have only partly overlapping valid times (fiscal years differ from calendar years). Concerning the metric Q_{31}, for all values, transaction time follows the end valid time.

5.3 Spatio-Temporal Assessment of Believability

Based on the metrics defined for each of the elementary sub-dimensions of believability, we perform the spatio-temporal assessment of the believability of data values. To achieve this, we aggregate metrics both along the different sub-dimensions of believability ("spatio") and along the lineage of data ("temporal"). We define the lineage of a data value v (noted Lineage(v)) as a labeled, directed acyclic graph representing the successive data values and processes leading to v. The data values are the vertices of the graph (noted V(Lineage(v))), and the processes are the labeled edges. In our case, we want to compute the spatio-temporal believability of data value v, and the lineage of v is represented in the bottom left part of Figure 3.

Our approach for computing spatio-temporal believability is developed in [13]. Due to space limitation, we only summarize (and simplify) the main steps of this approach, and apply it to the scenario of linked data used in the present paper.

The basic idea of our approach is to assess the spatio-temporal believability of a data value v by:

1. Aggregating believability metrics along the lineage of v, for each elementary sub-dimension of believability ("temporal" aggregation).
2. Aggregating the result along the different sub-dimensions of believability ("spatial" aggregation).

In order to aggregate believability metrics along the lineage of a data value v, we define an influence vector G(v), derived from the matrix M(v) as formalized below:

From the graph Lineage(v), build the matrix M(v) defined as follows:

*M(v) is a square matrix of order Card(V(Lineage(v))) * Card(V(Lineage(v)))*

The rows/columns of M(v) represent the vertices of Lineage(v). (The last row/column represents v.)

Call v_r *and* v_c *the vertices corresponding to row r and column c respectively.*

The content of element $M_{r,c}(v)$*, where r and c* $\in$ *1..Card(V(Lineage(v))), is defined as follows:*

If $\exists$ *an edge e from* v_r *to* v_c *in Lineage(v)*

Then Call *P the label of e.*

$$M_{r,c}(v) = \left| \frac{dP}{dx_r}(v_r) * v_r \right|$$

Else $M_{r,c}(v)=0$

Endif

For each *column c, divide each element* $M_{r,c}(v)$ *by the sum of the elements of column c.*

Let $N(v) = \sum_{k=0}^{\text{length(Lineage(v))}} (\gamma * \mathrm{M(v)})^k$

($\gamma \in [0..1[$ is an attenuation factor; length(Lineage(v)) is the length of the longest process chain from a source data value to v; the first term of the sum $\sum$, i.e. for k=0, is the identity matrix).

The final vector G(v) is the transpose of the last column of N(v), divided by the sum of elements of this column in order to normalize weights. (U is the unit column vector).

O(v)=

$(N(v)\,[1..\,Card(V(Lineage(v)));\ Card(V(Lineage(v)))])^T$

$G(v)=(1/(O(v) * U)) * O(v)$

Figure 5 shows the resulting matrices M(v) and G(v), for our example scenario. We assume a default value of 0.5 for the attenuation factor γ.

As an illustration of the computation of the values in Figure 5, consider the values of $M_{1,3}(v)$ and $M_{2,3}(v)$ (i.e. the values noted as a and b, respectively, in Figure 5). From Figure 3, we get the values for v_{11} and v_{22}, and we know that $P_1(v_{11},v_{12})=v_{11}+v_{12}$.

We have:

$$\mathrm{M}_{1,3}(\mathrm{v}) = \left|\frac{dP_1}{dx_{11}}(v_{11}) * v_{11}\right| / \left(\left|\frac{dP_1}{dx_{11}}(v_{11}) * v_{11}\right| + \left|\frac{dP_1}{dx_{12}}(v_{12}) * v_{12}\right|\right)$$

and

$$\mathrm{M}_{2,3}(\mathrm{v}) = \left|\frac{dP_1}{dx_{12}}(v_{12}) * v_{12}\right| / \left(\left|\frac{dP_1}{dx_{11}}(v_{11}) * v_{11}\right| + \left|\frac{dP_1}{dx_{12}}(v_{12}) * v_{12}\right|\right)$$

The derivative $\frac{dP_1}{dx_{11}}(v_{11}) = 1$ and, similarly, $\frac{dP_1}{dx_{12}}(v_{12}) = 1$. It follows that $\mathrm{M}_{1,3}$(v) = (1*109900000000) / (1*109900000000 + 1*203600000000) = 0.35 and $\mathrm{M}_{2,3}$(v) = (1*203600000000) / (1*109900000000 + 1*203600000000) = 0.65

Once we have the influence vector G(v), we may aggregate the believability metrics along the lineage of v, for each of the sub-dimensions of believability. In our example, we only consider the metrics Q_{31} and Q_{32}, as shown in Table 3. For Q_{31}, we perform a weighted average of the values shown in Table 3, using the weights as defined in the influence vector (weights shown in the right part of Figure 5). The result is 0.950. This number assesses Q_{31} (degree transaction follows valid time) by considering not only the value v, but also its lineage. For Q_{32}, we proceed similarly to perform aggregation along the lineage of v. We get the result 0.474 (if we consider null values as 1 for this metric), or 0.286 (if, in the weighted average, we consider only the defined values, i.e. the values for v_{21} and v). This figure (0.286) is higher than the non-aggregated value (0.167), reflecting the fact that v's lineage performs better than v itself on the sub-dimension "valid times overlap".

Finally, to perform the spatio-temporal believability of the data value v, we aggregate, along the different sub-dimensions of believability, the results obtained previously. We may choose between different aggregation operators [13]. In our example, we may use the product as aggregation operator. The final result for the believability of data value v (considering only two sub-dimensions of believability), is 0.950*0.286=0.272. This believability measure can be used, for example, to choose between alternative data sources (i.e. compute the data value v from other data sources and see how this choice affects the believability of v).

6. DISCUSSION AND CONCLUSION

We have presented a provenance model to capture and represent provenance information for computing data believability, and have illustrated how the provenance model may be operationalized in this context by mapping this model into RDF.

We have described our approach for assessing the believability of data, based on information stored in the provenance model. Data believability is composed of several sub-dimensions, for which specific metrics are defined. In this paper, we have focused on the metrics pertaining to the temporality of data, refining the metrics defined in previous work. Using an example scenario with linked data, we have illustrated how the believability of a data value is computed by aggregation along the lineage of the data value, and the sub-dimensions of believability.

This paper follows from previous research described in [13,14]. Compared with this previous work, the present paper (1) significantly extends and clarifies our temporality metrics and (2) operationalizes the approach in the Web of linked data. In order to operationalize the approach in the Web of linked data, we have mapped our provenance model into RDF, and applied our approach to an example of linked data, using data from data.gov.uk.

Instead of defining yet another RDF vocabulary for representing provenance in the semantic Web, we have chosen to use extant vocabularies. Mapping between our provenance model and RDF vocabularies and languages shows that almost all the concepts of our provenance model may be represented in the Web of linked data, although it is necessary to combine several vocabularies and languages.

We note that a provenance vocabulary for the Web of linked data is currently being defined (http://trdf.sourceforege.net/provenance/ns.html). However, the definition of this vocabulary is still underway, and the vocabulary is not as much used as other common vocabularies like FOAF for example. We also note that this vocabulary lacks some concepts important in our approach, e.g. valid time and trustworthiness concepts.

This work currently has some limitations. One limitation is that the Web of linked data is not much standardized yet in terms of vocabularies. Vocabularies generally differ from one data set to the other. Consequently, few data are readily available for automated, provenance-based computation of believability. Another limitation concerns the determination of valid time. In our approach, we are primarily interested in assessing the believability of decision-oriented (as opposed to transactional) data, e.g. statistics. For this kind of data, the valid time is generally known (for example, there is normally a time dimension in an OLAP cube). But the problem in the Web of linked data is that valid time is not always represented in an atomic, machine-readable form as is the case with the data.gov.uk data set.

Further work will consist in testing our approach with other data sets, and further refining the metrics defined for the believability sub-dimensions of data quality.

7. REFERENCES

[1] Agrawal, R. et al. The Claremont report on database research. Comm. ACM 52, 6 (June 2009), 56-65.

[2] Ballou, D., and Pazer, H. Modeling data and process quality in multi-input, multi-output information systems. Manage. Sci. 31, 2 (February 1985), 150-162.

[3] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. Methodologies for data quality assessment and improvement. ACM Comput. Surv. 41, 3 (July 2009), article 16.

[4] Cohen-Boulakia, S., Biton, O., Cohen, S., and Davidson, S. Addressing the provenance challenge using ZOOM. Working Paper, Department of Computer & Information Science, University of Pennsylvania, May 2007, http://repository.upenn.edu/cis_papers/338/

[5] Dubois, D., and Prade, H. An introduction to possibilistic and fuzzy logics. In Smets, P., Mamdani, A., Dubois, D., and Prade, H. (eds), Non-standard logics for automated reasoning. Academic Press, London, 1988.

[6] Even, A., and Shankaranarayanan, G. Utility-driven assessment of data quality. Data Base 38, 2 (May 2007), 75-93.

[7] Forrester, J. Industrial dynamics-after the first decade. Manage. Sci. 14, 7 (March 1968), 398-415.

[8] Hartig, O. Provenance information in the Web of data. Proceedings of LDOW 2009 (Madrid, Spain, April 2009).

[9] Jensen, C., et al. A consensus glossary of temporal database concepts. SIGMOD Record 23, 1 (March 1994), 52-64.

[10] Lee, Y., Pipino, L., Funk, J., and Wang, R. Journey to Data Quality. MIT Press, Cambridge, MA, 2006.

[11] Moreau, L. et al. The provenance of electronic data. Comm. ACM 51, 4 (April 2008), 52-58.

[12] Object Management Group, Unified Modeling Language (UML) specification, version 2.2, http://www.omg.org/technology/documents/formal/uml.htm

[13] Prat, N., and Madnick, S. Evaluating and aggregating data believability across quality sub-dimensions and data lineage. Proceedings of WITS 2007 (Montreal, Canada, December 2007), 169-174.

[14] Prat, N., and Madnick, S. Measuring data believability: a provenance approach. Proceedings of HICSS-41 (Big Island, HI, January 2008), IEEE, 1-10.

[15] Ram, S., and Liu, J. Understanding the semantics of data provenance to support active conceptual modeling. Proceedings of ACM-L 2006 (Tucson, AZ, November 2006), 17-29.

[16] Shadbolt, N., Hall, W., and Berners-Lee, T. The semantic Web revisited. IEEE Intell. Syst. 21, 3 (May 2006), 96-101.

[17] Simmhan, Y.L., Plale, B., and Gannon, D. A survey of data provenance in e-science. SIGMOD Record 34, 3 (September 2005), 31-36.

[18] Tseng, S., and Fogg, B. Credibility and computing technology. Comm. ACM 42, 5 (May 1999), 39-44.

[19] Wang, R, and Strong, D. Beyond accuracy: what data quality means to data consumers. J. Manage. Inform. Syst. 12, 4 (Spring 1996), 5-34.

[20] Widom, J. Trio: a system for integrated management of data, accuracy, and lineage. Proceedings of CIDR 2005 (Asilomar, CA, January 2005), 262-276.

APPENDIX: USE OF THE PROVENANCE MODEL IN BELIEVABILITY METRICS

Concept of the provenance model	**Believability metrics**
Person	
• groups	trustworthiness of source
• organizations	trustworthiness of source
Group	
• organization	trustworthiness of source
Domain	
• sub_domain_of	trustworthiness of source
Trustworthiness	
• agent	trustworthiness of source
• domain	trustworthiness of source
• trustworthiness value	trustworthiness of source
Data	
• data_label	consistency over sources, consistency over time
• data_type	
• is_stock	valid times overlap
• start_valid_time	valid times overlap, consistency over time
• end_valid_time	valid times overlap, degree transaction follows valid time , consistency over time
Possibility_distribution	
• possibility_distribution	possibility
Data_value	
• data_value	trustworthiness of source, possibility, consistency over sources, consistency over time
• data	possibility, consistency over sources, consistency over time
Source_data_value	
• provided_by	trustworthiness of source, consistency over sources
• transaction_time	degree transaction follows valid time
Resulting_data_value	
• process_run	trustworthiness of source
Process_run	
• under_responsibility_of	trustworthiness of source
• transaction_time	degree transaction follows valid time
• process	trustworthiness of source
Process	
• input	trustworthiness of source
• output	trustworthiness of source