**The International Journal of Digital Curation**

**Volume 7, Issue 1 | 2012**

# Requirements for Provenance on the Web

Paul Groth,
VU University, Amsterdam

Yolanda Gil,
Information Sciences Institute,
University of Southern California

James Cheney,
University of Edinburgh

Simon Miles,
Kings College London

## Abstract

From where did this tweet originate? Was this quote from the New York Times modified? Daily, we rely on data from the Web, but often it is difficult or impossible to determine where it came from or how it was produced. This lack of provenance is particularly evident when people and systems deal with Web information or with any environment where information comes from sources of varying quality. Provenance is not captured pervasively in information systems. There are major technical, social, and economic impediments that stand in the way of using provenance effectively. This paper synthesizes requirements for provenance on the Web for a number of dimensions, focusing on three key aspects of provenance: the content of provenance, the management of provenance records, and the uses of provenance information. To illustrate these requirements, we use three synthesized scenarios that encompass provenance problems faced by Web users today.

*International Journal of Digital Curation* (2012), *7*(1), 39–56. http://dx.doi.org/10.2218/ijdc.v7i1.213

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by UKOLN at the University of Bath and is a publication of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

# Introduction

With the availability of massive amounts of data (Berman, 2008; Bizer et al., 2009), information about the provenance of that data becomes an important factor in developing new applications. The provenance of information is crucial to making determinations about whether information is trusted, how to integrate diverse information sources, and how to give credit to originators when reusing information (Cheney et al., 2009; Simmhan et al., 2005; Buneman et al., 2008). Broadly construed, provenance encompasses the initial sources of information used, as well as any entity and process involved in producing a result (Moreau, 2010). Provenance is useful in many contexts, but more so in an open and inclusive environment, such as the Web, containing information that is often contradictory or questionable. People make trust judgments based on provenance that may or may not be explicitly offered to them (Artz & Gil, 2007). Therefore, a crucial enabler of automation on the Web is the explicit representation of provenance that is accessible to machines, not just to humans.

Even recognizing that provenance is central to the way that humans make use of information, typically information systems offer little or no support for provenance beyond primitive (and unreliable) ownership, creation and modification timestamps. We are not routinely exposed to software, databases, or Web applications that understand and handle rich forms of provenance for us. However, we do use provenance in all our decision making. Why is provenance not pervasive in all information systems and software? We suspect that although the treatment of provenance seems straightforward (after all there is nothing conceptually challenging about recording and replaying extra log information), the design of appropriate provenance solutions becomes complex very quickly and that these complexities are under-appreciated by the research community at large, as well as by practitioners and system developers.

Provenance may be in an analogous situation to how user interfaces were treated a few decades ago. That is, user interfaces were viewed as an afterthought to a system's design (after all there is nothing challenging about putting a few buttons and menus here and there) but then in practice they became complex very quickly. Appropriate methodology had to be developed in order to understand the role of a user interface in software systems. Today, user interfaces are at the forefront of a system's design, ensuring the usability and success of the system. Provenance may be in the early stages of a similar cycle. We lack the principles, languages and methodologies to incorporate provenance in the design and implementation of software systems more pervasively and thoroughly. Provenance solutions may boost the usability and ultimately the success of today's software systems, and perhaps open the door to new application areas where delegation and trust are paramount. The aim of this paper is to outline the requirements that arise in designing provenance systems and therefore uncover the difficulties involved.

Although this paper has broad relevance, our focus is provenance on the Web as an open information system. By considering Web-accessible data as the subject of provenance, the emphasis is on contexts where the user of data may have very little connection with its providers or curators, and where the provenance will often not describe a ‘closed’ process within one institution, but a set of interconnected processes in which we repeatedly find data from one source being used to derive data by a distinct and independent user. Driven by the increasing need to manage the diversity and varying quality of Web data and information, the W3C chartered the Provenance Incubator Group[1] to provide state-of-the art understanding and develop a roadmap in the area of provenance for Semantic Web technologies, development, and possible standardization. The group produced a number of documents, including a report of key dimensions for provenance, more than thirty use cases spanning many areas and contexts that illustrate these key dimensions, and a broad set of user requirements and technical requirements derived from those use cases. The group’s reports are available in its public Web site.[2] The incubator group successfully proposed a standardization effort regarding provenance, drawing on the incubator group’s results, and this W3C working group is ongoing at the time of writing.[3]

This paper summarizes the group’s findings regarding user requirements for provenance, and provides an analysis of the aspects, or dimensions, which any general provenance solution for Web data must consider. By connecting these dimensions to concrete use cases, we help make the meaning of those dimensions more explicit and explicable. Further, we categorize the dimensions to enable a separation of concerns in engineering solutions: those of concern when deciding how to model provenance content; those of concern when storing, maintaining and providing access to provenance data; and those of concern for allowing users to achieve what they want. This paper does not present solutions to these requirements but instead aims to scope and define the problem of provenance in the Web space.

First, we describe three scenarios that were designed to cover different subsets of requirements and address different communities of interest. After describing the three scenarios, we discuss the requirements for provenance.

## Motivating Scenarios: The Need for Provenance

We begin by describing three broad scenarios that illustrate the need for provenance. They are based on over 30 original use cases collected by the W3C Provenance Group.

### News Aggregator Scenario

Many Web users would like to have mechanisms to automatically determine whether a Web document or resource can be used based on the original source of the content, the licensing information associated with the resource, and any usage restrictions on that content. Furthermore, in cases of mashed-up content, it would be useful to ascertain automatically whether or not to trust it by examining the processes that created, aggregated, and delivered it, as well as who was responsible for these processes. To illustrate these issues, we present the following scenario of a fictitious Web site, BlogAgg, that aggregates news information and opinion from the Web.

[1] W3C Provenance Incubator Group: http://www.w3.org/2005/Incubator/prov/charter

[2] W3C Provenance Incubator Group Wiki: http://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki

[3] Provenance Working Group: http://www.w3.org/2011/prov/

BlogAgg aims to provide rich, real time news to its readers automatically. It does this by aggregating information from a wider variety of sources, including microblogging Web sites, news Web sites, publicly available blogs and other opinion pieces. It is imperative for BlogAgg to present only credible and trusted content on its site to satisfy its customers, attract new business, and avoid legal issues. Importantly, it wants to ensure that the news items that it aggregates are correctly attributed to the right person so that they may receive credit. Additionally, it wants to present the most attractive content it can find, including images and video. However, unlike other aggregators, it wants to track many different Web sites to try and find the most up to the date news and information.

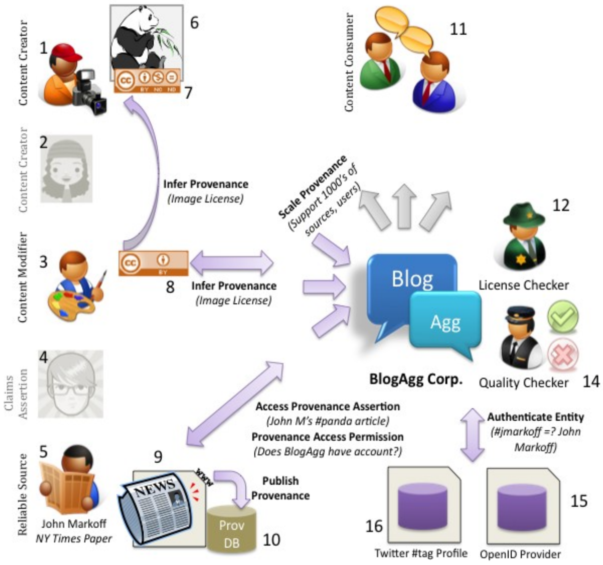

Figure 1. Overview of provenance issues in the News Aggregator Scenario

Unfortunately for BlogAgg, the source of the information is not often apparent from the data that it aggregates from the Web. In particular, it must employ teams of people to check that selected content is both high-quality and can be used legally. The site would like this quality control process to be handled automatically.

Consider a hypothetical scenario, illustrated in Figure 1, where one day BlogAgg discovers that #panda is a trending topic on Twitter. It finds that the tweet "*#panda being moved from Chicago Zoo to Florida! Stop it from sweating http://bit.ly/ssurl*" is being retweeted across many different microblogging sites. BlogAgg wants to find the correct originator of the microblog who first got the word out. It would like to check if it is a trustworthy source (whether organization or person) and verify the news story. It would also like to credit the original source in its site, and in these credits it would like to include some identifying information about that person, for example, a link to a Facebook profile. In cases where the person wishes to remain anonymous, BlogAgg would like to provide some sort of information that shows the source is authoritative and/or trustworthy on the topic.

Following the tiny-url, BlogAgg discovers a site protesting about the move of the panda. BlogAgg wants to determine what author (and their affiliation) is responsible for the site so that its name can run next to the snippet of text that BlogAgg runs. In determining the snippet of text to use, BlogAgg needs to determine whether or not the text originated at the panda protest site or was a quoted from another site. Additionally, the site contains an image of a panda that appears as if it is sweating. BlogAgg would like to automatically use a thumbnail version of this image in its site, therefore, it needs to determine if the image license allows this or if that license has expired and is no longer in force. Furthermore, BlogAgg would like to determine if the image was modified, by whom, and if the underlying image can be reused (i.e. whether the license of the image is actually correct). Additionally, it wants to find out whether any modifications were just touch-ups or were significant modifications. Using the various information about the content it has retrieved, BlogAgg creates an aggregated post. For the post, it provides a visual seal showing how much the site trusts the information. By clicking on the seal, the user can inspect how the post was constructed and from what sources and a description of how the trust rating was derived from those sources.

After publication of the panda piece by BlogAgg, the original author of the panda tweet discovers that the information that they had was incorrect and wants to retract the statement. She therefore removes (or clicks the retract button) on Twitter. BlogAgg periodically checks to see if any of the information it used to produce a piece has been retracted. If it has, it updates the post to reflect this retraction. During these checks, BlogAgg also checks whether their information is up-to-date. For example, if it's pointing at the correct identity for the author, whether that information has moved or whether it has been deleted.

BlogAgg wants to repeat this process for thousands to hundreds of thousands of sites a day. It wants to automate this aggregation and checking process as much as possible. It is important for BlogAgg to be able to detect when this aggregation process does not work, in particular when it cannot determine the origins of the content it uses. These need to be flagged and reported to BlogAgg's operators.

### Disease Outbreak Scenario

Many uses of the Web involve the combination of data from diverse sources. Data can only be meaningfully reused if the collection processes are exposed to users. This enables the assessment of the context in which the data was created, its quality and validity, and the appropriate conditions for use. Often data is reused across domains of interest that end up mutually influencing each other. This scenario, illustrated in Figure 2, focuses on the reuse of data across disciplines in both anticipated and unanticipated ways.

Alice is an epidemiologist studying the spread of a new disease called owl flu (a fictitious disease made up for this example), with support from a government grant. Many studies relevant to public policy are funded by the government, with the expectation that the conclusions will provide guidance for public policy decision-makers. These decisions need to be justified by a cost-benefit analysis. In practice, this means that results of studies not only need to be scientifically valid, but the source data, intermediate steps and conclusions all need to be available for other scientists or non-experts to evaluate. In the United Kingdom, for example, there are published guidelines that scientists are required to follow in their reports (HM Treasury, 2003).

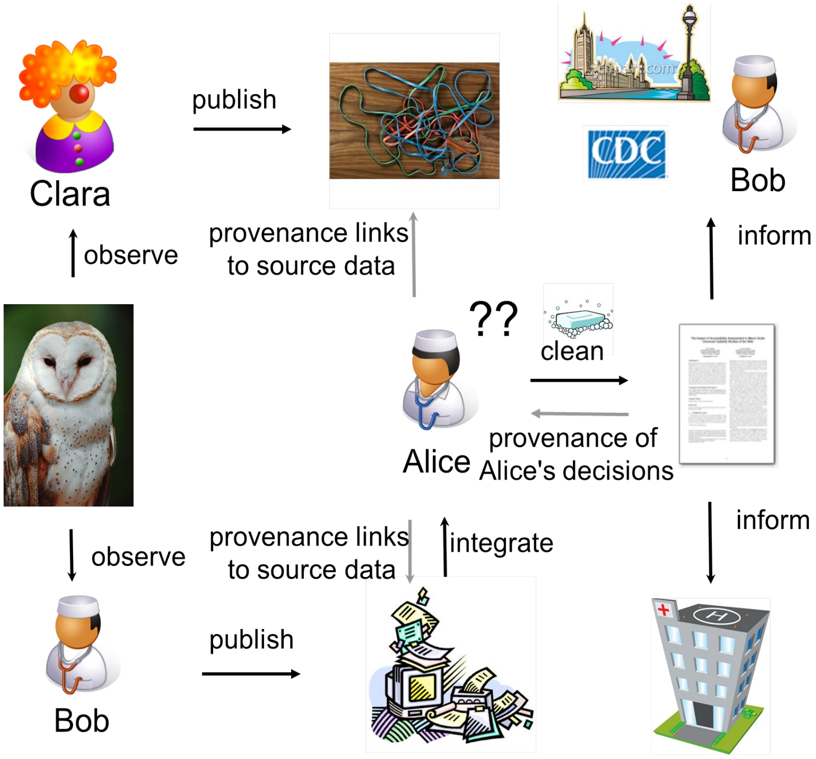
Clara
publish
provenance links to source data
observe
CDC
Bob
inform
??
clean
Alice
provenance of Alice's decisions
inform
observe
provenance links to source data
integrate
Bob
publish

Figure 2. Overview of provenance issues in the Disease Outbreak Scenario.

Alice's study involves collecting reports from hospitals and other health services, as well as recruiting participants, distributing and collecting surveys, and performing telephone interviews to understand how the disease spreads. Some of the data may be available from hospital report sites, and may contain details about how they were collected and organized. Some of the data collected by Alice herself may be initially recorded on paper and then transcribed into an electronic form. The paper records are confidential, but need to be retained for a set period of time.

Alice will also use data from public sources, such as Data.gov[4], or repositories, such as the UK Data Archive.[5] Currently, a number of e-Social Science archives (such as NeISS, e-Stat and NESSTAR) are being developed for keeping and adding value to such datasets, which Alice may also use. Alice may also seek out "natural experiment" data sources that were originally gathered for some other purpose, such as customer sales records, tweets on a given day, or free geospatial data.

Once the data are collected and transcribed, Alice processes and interprets the results and writes a report summarizing the conclusions. Processing the data is labor-intensive, and may involve using a combination of many tools ranging from spreadsheets, generic statistics packages (such as SPSS or Stata), or analysis packages that cater specifically to Alice's research area.

[4] Data.gov: http://data.gov.uk

[5] UK Data Archive: http://data-archive.ac.uk

Alice may find many challenges in integrating the data from different sources, since units or semantics of fields are not always documented. When different data sources use different representations, Alice may need to recode or integrate this data by hand or by guessing an appropriate conversion function. These are subjective choices that may need to be revisited later. To prepare the final report, Alice may also make use of advanced visualization tools, for example to create a Web mashup that plots results on maps.

The conclusions of the report may then be incorporated into policy briefing documents used by civil servants or experts on behalf of the government to identify possible policy decisions that will be made by (non-expert) decision makers, to help avoid or respond to future outbreaks of owl flu. This process may involve considering hypotheticals or re-evaluating the primary data using different methodologies than those applied originally by Alice. The report and its linked supporting data may also be published on the Web for reuse by others or archived permanently in order to compare the predictions made by the study with the actual effects of decisions in the future. In some cases, the data used by Alice may not be available on the open Web. Thus, it may be impossible to reuse the underlying data. Furthermore, the identifiers (e.g. URNs) may or may not be available depending on the status of the information.

Bob is a biologist working to develop diagnostic and chemotherapeutic targets in the human pathogen responsible for owl flu. Bob’s experimental data may be combined with data produced in Alice’s epidemiological study, and Bob’s level of trust in this data will be influenced by the detail and consistency of the provenance information accompanying Alice’s published report on the Web. Bob generates data using different experiment protocols, such as expression profiling, proteome analysis, and creation of new strains of pathogens through gene knockout. These experiment datasets are also combined with data from external sources available over the Web, such as curated biology databases (NCBI Entrez Gene/GEO, TriTrypDB, EBI’s ArrayExpress) and information from biomedical literature (PubMed) that have different curation methods and quality associated with them. Biologists need to judge the quality, timeliness and relevance of these sources, particularly when data needs to be integrated or combined.

These sources are used to perform “in silico” experiments via scientific workflows or other programming techniques. These experiments typically involve running several computational steps, such as running machine learning algorithms to cluster gene expression patterns or to classify patients based on genotype and phenotype information. The results need to meet a high standard of evidence so that the biologists can publish them in peer-reviewed journals. Therefore, justification information documenting the process used to derive the results is required. This information can be used to validate the results, to understand unexpected results, to integrate results, and to infer heuristics for data quality. All this process documentation may be published on Bob’s project Web site in a repository of detailed records that his collaborators can access and query as Web objects.

As more data of owl flu outbreaks and treatments become available over time, Alice’s epidemiological studies and Bob’s research on the behavior of the owl flu virus will need to be updated by repeating the same analytical processes incorporating the new data.

## Business Contract Scenario

In scientific collaborations and in business, individual entities often enter into some form of contract to provide specific services and/or to follow certain procedures as part of the overall effort. Proof that work was performed in conformance with the expectations of the project leadership (as expressed in the contract) is often required in order to receive payment or to settle disputes. Such proof must, for example, document work that was performed on specific items, provide evidence that would preclude various types of fraud, allow a combination of evidence from multiple witnesses, and be able to provide partial information to protect privacy or trade secrets. To illustrate such demands of proof, and the other requirements which stem from having such information, we consider the following use case, illustrated in Figure 3.

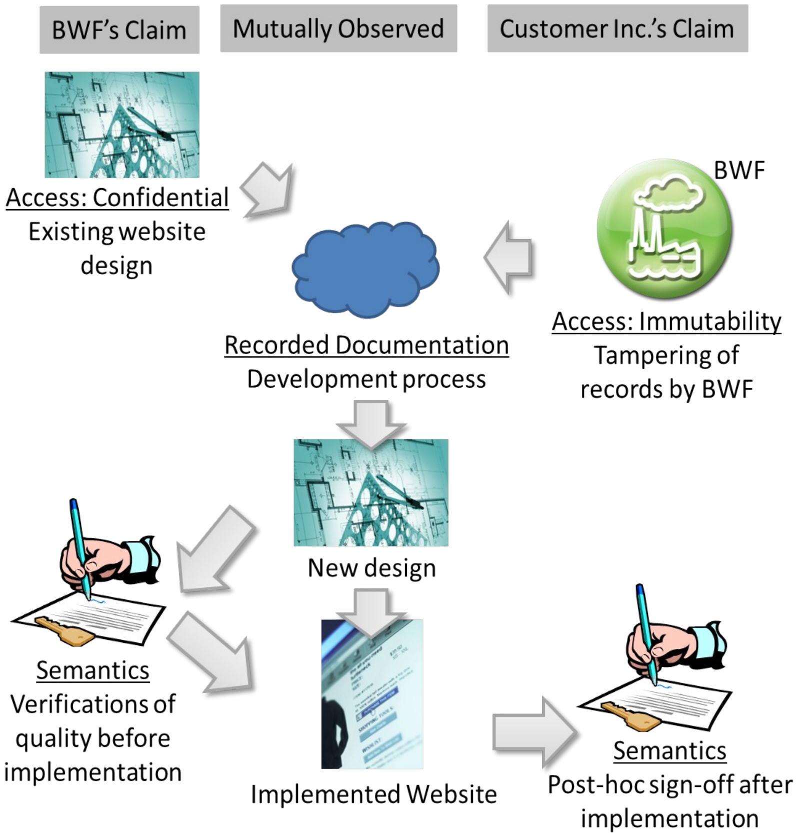

Figure 3. Overview of provenance issues in the Business Contract Scenario.

Bob's Web site Factory (BWF) is a fictitious company that creates Web sites which include secured functionality, such as for payments. Customers choose a template Web site structure, enter specifications according to a questionnaire, and upload company graphics. BWF will then create an attractive and distinct Web site. The final stage of purchasing a Web site is for the customer to agree to a contract setting out the responsibilities of the customer and BWF. The final contract document, including BWF's digital signature, will be automatically sent by email to both parties.

BWF has completed a contract with a customer, Customers Inc., for the supply of a Web site. Customers Inc. are not happy with the product they receive, and assert that contractual quality requirements were not met. Customers Inc. claims that BWF failed to create a site to their specifications, that the work was performed by someone without adequate training, and that the security of payments to the site is faulty due to improper quality control and testing procedures not being followed. Finally, Customers Inc. claim that records were tampered with to remove evidence of the latter improper practices. BWF wishes to defend itself against these claims.

BWF will need to provide proof that the contract was executed as agreed. However, they have concerns about what information to include in such a proof. Many Web sites are not designed from scratch but are based on an existing design in response to the customer's requests or problems. Also, sometimes parts of the design of multiple different sites, designed for other customers, are combined to make a new Web site. To protect its own intellectual property and confidential information regarding other customers, BWF wishes to reveal only that information needed to defend against Customers Inc.'s claims.

There are many kinds of entities relevant to describing BWF's development processes, from source code text in programs to images. To provide adequate proof, BWF may need to include documentation on all of these different objects, making it possible to follow the development of a final design through multiple stages. The contract number of a site is not enough to unambiguously identify that artifact, as designs move through multiple versions: the proof requires showing that a site was not developed from a previous version of a flawed design or using security software known to be faulty.

To show that there was adequate quality control and testing of the newly designed or redesigned sites, BWF needs to demonstrate that a design was approved in independent checks by at least two experts. In particular, the records should show that the experts were truly independent in their assessments, i.e. they did not both base their judgement on one, possibly erroneous, summarized report of the design.

Finally, Customers Inc. claim that there is a discrepancy in the records, suggesting tampering by BWF to hide their incompetence, as the development division apparently claimed to have received instructions from the experts checking the design that it was OK before the experts themselves claim to have supplied such a report. BWF suspects, and wishes to check, that this is due to a difference in semantics between the reported times, e.g. in one case it regards the receipt of the report by the developers, in the other it regards the receipt of the acknowledgement of the report from the developers by the experts. These reports should be shared in a format that both parties understand.

## Requirements for Provenance on the Web

While the exact requirements for a system enabling provenance capture and retrieval will depend on the scope of envisaged use cases and the technology upon which it is based, we can make explicit the dimensions that the requirements will fall under. We can further support engineers of those systems by categorizing the dimensions according to the kind of functionality they imply. Specifically, those modelling what should be contained in provenance data (its content) will have different concerns from those developing software to capture and maintain the data (its management), and again different concerns from those providing solutions to solve specific user problems (its use). Table 1 provides a description of the dimensions and their categories. We now illustrate this categorization using the scenarios described above.

| Category | Dimension | Description |
|---|---|---|
| Content | Object | The artifact that a provenance statement is about. |
| | Attribution | The sources or entities that contributed to create the artifact in question. |
| | Process | The activities (or steps) that were carried out to generate the artifact at hand. |
| | Versioning | Records of changes to or between artifacts over time, and what entities and processes were associated with those changes. |
| | Justification | Documentation recording why and how a particular decision is made. |
| | Entailment | Explanations showing how facts were derived from other facts. |
| Management | Publication | Making provenance available on the Web. |
| | Access | The ability to find the provenance for a particular artifact. |
| | Dissemination | Defining how provenance should be distributed and controlled. |
| | Scale | Dealing with large amounts of provenance. |
| Use | Understanding | How to enable the end user consumption of provenance. |
| | Interoperability | Combining provenance produced by multiple different systems. |
| | Comparison | Comparing artifacts through their provenance. |
| | Accountability | Using provenance to assign credit or blame. |
| | Trust | Using provenance to make trust judgments. |
| | Imperfections | Dealing with imperfections in provenance records. |
| | Debugging | Using provenance to detect failures or bugs. |

Table 1. Major provenance dimensions.

Note that the dimensions listed are not intended to be a provenance vocabulary themselves, but are aspects that we argue provenance models, existing or to be developed, should address in order to enable the common use cases for provenance on the Web. As such, they can be seen as a framework for assessing the adequacy of those models or their supporting software systems.

## Content

The Content dimension refers to the structure and meaning of provenance records. Although it appears straightforward to design a data model for provenance sufficient for a fixed application, the challenge is to standardize a format that is suitable for a broad range of uses.

We need to establish first what the artifact that is the **object** of any statements about provenance is, and be able to refer to it. On the Web this will be a Web resource, essentially anything that can be identified with a URI, such as Web documents, datasets, assertions, or services. In the News Aggregator Scenario, the object of provenance was the final news item posted by the news aggregator. It can also be a set or collection of items, such as the pages of the Web site sold in the Business Contract Scenario. Conversely, sometimes provenance information needs to refer to a particular portion or aspect of an artifact. It is especially challenging to keep track of provenance during an object's lifetime as it migrates among different systems. For example, objects may be organized in collections, then subgroups selected, then portions of some objects modified, then the document may be disseminated on the Web and its copies further modified.

One important aspect of content is **Attribution**, which of the sources (i.e., typically any Web resource that has an associated URI, such as documents, Web sites, or data) or entities (i.e., people, organizations, and other identifiable groups) that contributed to create the artifact in question. In the News Aggregator Scenario, sources would be the URIs for a blog or for a news item at a Web site, while entities would include Twitter.com and the person who created the original microblog. Typically, an artifact would be associated with several sources and entities. One challenge in this respect is to represent which of them has **responsibility** for the artifact at hand, meaning which of the many entities that were associated with its creation actually endorses or stands by it. For example, in the News Aggregator Scenario the image of the panda may have been modified before it was posted on the protest site but some entity must have originally taken the actual picture of the unhappy panda and endorsed it. In addition, the provenance representation of attribution should also enable us to see the true **origin** of any statement of attribution to an entity. It is important to represent whether the statement was recorded by that entity and can be verified (for example, with a digital signature). Alternatively, we would want an indication that the attribution statement was stated by the original entity or was reconstructed and then asserted by a third party.

Another important aspect of provenance content is **process**. This refers to the activities (or steps) that were carried out to generate the artifact at hand. These activities encompass the execution of a computer program that we can explicitly point to, a physical act that we can only refer to, and some action performed by a person that can only be partially represented. An example of such activity in the Disease Outbreak Scenario is the execution of a machine learning algorithm to create a classification for a clinical trial dataset. Provenance information may need to refer to descriptions of the activities, so that it becomes possible to support reasoning about processes involved in provenance and support descriptive queries about them. For example, in the Disease Outbreak Scenario, the machine learning algorithm used may be described as a Bayes algorithm or a decision tree algorithm, and a user may query for whether the classifier is human readable or not (which the latter is while the former is not). Processes can be represented at a very abstract level, focusing only on important aspects, or at a very fine-grained level to include minute details. For example, in the Disease Outbreak Scenario, users may like to know how a query result was computed. More specifically, users may like to know the data elements and query language operators that were used to compute the result of a query.

With respect to the detail of provenance content, an important consideration is to represent enough process details to allow **reproducibility**; that is, the ability to re-create the artifact by re-enacting (re-executing) the process described by its provenance record. For example, in the Disease Outbreak Scenario, a scientist may want to re-run the same machine learning algorithm to include more recent data and check whether they obtain the same classification or a different one. For this they would need to have access to the code and perhaps the parameter settings, while it may not be important to know which machine was used. Reproducibility may only be possible for some period of time after publication of provenance, as it is likely that the codes or data may not be available after some time. The shelf life of reproducible mechanisms could be stated as part of the provenance. After that time, at least inspectability should be possible, where the details of what was executed can be examined even if the process cannot be re-executed. Process provenance information needs to be connected with attribution information, so that it is possible to represent how sources or entities were involved in what activities and what their role was. For example, in the Disease Outbreak Scenario, the clinical dataset was a source that played the role of being input to the machine learning algorithm, while the scientist was associated with that same activity as the entity that set its parameters. **Resource access** is an important aspect of representing process provenance. This includes matters such as the access time, the server accessed, and the party responsible for the server.

**Evolution and versioning** must be treated with special status in a provenance representation. When one has full control over an artifact and its provenance records this may be a simple matter of housekeeping, but this is a challenge in open distributed environments, such as the Web. Consider the representation of provenance when **republishing**, for example by retweeting, reblogging, or repackaging a document. It should be possible to represent when a set of changes constitute a new version of the object in question. In addition, it is an open question whether each new version published should publish not only the artifact, but also its full provenance, as available in the original source. Another important issue is the tension between the completeness of provenance and privacy and confidentiality concerns. In the Business Contract Scenario, when the design of a Web site developed for one customer has aspects drawn from designs for other customers, publishing the full provenance may breach confidentiality: information might be revealed about past customers. Another open question is how to associate provenance information to each of the **updates** to an artifact. One must consider whether the entire provenance trail should be associated with each update, or whether each change should only represent the delta differences from prior versions of the artifact, which can then be consulted for more extensive provenance. Dealing with versioning is difficult on the Web, as what is identified by a URI (e.g. a Web resource) may change. In addition, whether a resource has changed version can often be difficult to understand because the representations of resources may differ but the underlying resource may be constant. Thus, versioning becomes not only a problem at the level of the data being considered but also of the Web architecture.

A particular kind of provenance information is **justifications of decisions**. The purpose of a justification is to allow those decisions to be discussed and understood. In the Disease Outbreak Scenario, it is important to capture the arguments for and against the conclusions of the owl flu report, as well as the ability to capture the evidence behind particular hypotheses in the report.

Some provenance information may be directly asserted by the relevant sources of some data or actors in a process, while other information may be derived from that which was asserted. In general, one fact may **entail** another, and this is important in the case of provenance data that is inherently describing the past, for which the majority of facts cannot now be known. In the example from the Business Contract Scenario, the complainants in the case derive from the (ambiguous) records held by the company that suggest that a check on the quality of security-related code was performed only after development based on that code had started. This derivation should not be taken as plain fact by the court, but understood as due to a particular derivation process, with a set of assumptions, and by a particular (biased) source.

### Management

The **management** dimension refers to the mechanisms that make provenance available and accessible in a system. Provenance management poses many substantial challenges to system design and implementation.

An important issue is the **publication** of provenance. Provenance information must be made available on the Web. Related issues include how provenance is exposed, discovered, and distributed. A **provenance representation language** must be chosen and made available so others can refer to it in interpreting the provenance. The **publisher** of provenance information should be associated with provenance records.

Once provenance is available, it must be **accessible**. That is, it must be possible to find it by specifying the artifact of interest. For example, in the News Aggregator Scenario, the aggregator must be able to point to a tweet and query about its provenance. **Query formulation and execution** mechanisms must be defined for provenance representation.

In realistic settings, provenance information will have to be subject to **dissemination control**. Provenance information may be associated with **access policies** about what aspects of provenance can be made available given a requestor's credentials, as illustrated in the Disease Outbreak Scenario. Provenance may have associated **use policies** about how an artifact can be used, given its origins. This may include **licensing** information stated by the artifact's creators regarding what rights of use are possible for the artifact. For example, in the News Aggregator Scenario, the aggregator needs to be able to determine what license the panda image actually has. Finally, provenance information may be withheld from access for **privacy protection**. In the Business Contract Scenario, the Website developer has a need to filter what provenance information is revealed about a design due to its entanglement with their own intellectual property and confidential information about other customers.

The **scale** of provenance information is a major concern, as the size of the provenance records may by far exceed the scale of the artifacts themselves. This poses significant data management and querying challenges, which databases and other systems are only beginning to address. Tradeoffs must be made regarding the granularity of the provenance records kept and the actual amount of detail needed by users of provenance. For instance, in the Disease Outbreak Scenario, the complete provenance about the result of an analysis may include a huge portion of the biomedical literature distributed across many published repositories.

### Use

The **use** dimension refers to the aims that motivate recording and managing provenance. The same provenance records may be expected to serve a variety of purposes, some of which may initially be ill-understood. We view identifying clear **specifications** and **policies** for provenance mechanisms as a key challenge.

An important consideration is how to make provenance information **understandable** to its users/consumers. Just because the information that they need is recorded in the provenance representation it does not guarantee that it can be used for their purposes. To this end, it would be useful to support multiple levels of **abstraction** in the provenance records of an artifact, as well as multiple perspectives or views over such provenance. In the Disease Outbreak Scenario, a scientist may want to start with a high-level description that includes only the machine learning algorithms used but not any details of the data format conversion operations that were carried out, and then the latter may be shown upon request. In addition, appropriate **presentation and visualization** of provenance information is an important consideration, as users will likely prefer something other than a set of provenance statements. For example, in the Disease Outbreak Scenario, a scientist may prefer to look at a dataflow diagram of the processes used to generate a result, while in the News Aggregator Scenario, a simple logo signifying approved provenance (with a link to more detailed explanation) may be more appropriate to use instead.

Because provenance information may be obtained from heterogeneous systems and different representations, **interoperability** is an important requirement. A query may be issued to retrieve information from provenance records created by different systems that then need to be integrated. At a finer grain, the provenance of a given artifact may be specified by multiple systems and need to be combined. It could be that each system contributed a type of provenance information, or that different systems contributed the provenance of sources used to create the artifact. Each news site of the News Aggregator Scenario may use a different representation of provenance, and the aggregator would need to integrate them to provide a coherent provenance picture to the user.

Another important use of provenance is for **comparison** of artifacts based on their origins. Two artifacts may seem very different while their provenance may indicate significant commonalities. Conversely, two artifacts may seem alike, and their provenance may reveal important differences. For example in the Disease Outbreak Scenario, two scientists may want to compare results from their experiments by comparing the process provenance information as well as the entities (e.g., reagents) that they each used.

Provenance data can be used for **accountability**, such as in the Business Contract Scenario, where the Web site engineering company uses provenance data to account for its actions. Accountability requires that the users can rely on the provenance record and authenticate its sources.

A very important use of provenance is **trust**. Provenance information is used to make trust judgments on a given entity. For example, in the News Aggregator Scenario, the aggregator site makes decisions about whether to include a news item based on its provenance. Users can make similar decisions based on provenance as well, for example, by defining filters to eliminate news items based on specific properties of their provenance information. Trust is often based on attribution information, by checking the **reputation** of the entities involved in the provenance, perhaps based on past reliability ratings, known authorities, or third-party recommendations. Similarly, measures of the relative **information quality** can be used to choose among competing evidence from diverse sources. For example, in the Disease Outbreak Scenario, researchers will need to assess the quality of the data they are using, particularly if they are obtained from open public sources that do not follow formal creation or curation processes. Finally, users should be able to access and understand how automated trust assessments are derived from provenance.

Using provenance information may imply handling **imperfections**. Provenance information may be **incomplete** in that some information may be missing, or **incorrect** if there are errors. Provenance information may also be provided with some **uncertainty** or be of a probabilistic nature. These imperfections may be caused because of problems with the recording of the provenance information. But they may also arise because the user does not have access to the complete and accurate provenance records, even if they exist. This is the case when provenance is summarized or compressed, in which case many details may be abstracted away or missing. Finally, provenance may also be subject to deception and be fraudulent partially or in its entirety. Several of the latter cases can be seen in the Business Contract Scenario: the engineering company hides some records due to confidentiality, leading to gaps in the account, while the complainant claims that this and the apparent time discrepancies suggest that the records have been fraudulently altered to support the engineering company’s case.

Another use of provenance is **debugging**. An example is shown in the Business Contract Scenario, where the company tries to determine whether it has made an error due to two apparently independent evaluations being dependent on the same faulty source information. Without a record of the provenance of those evaluations, it would be impossible to determine whether such a “bug” in the process had indeed occurred.

## Related Work

While provenance is not widely embedded in systems, the problem itself is not new. There is a wealth of literature on provenance that can be leveraged in addressing the requirements outlined in this paper.

In computer science, there are several overviews of technical research contributions regarding provenance. A comprehensive review is provided by Moreau (2010), which lays a foundation for the how these technologies can support provenance on the Web.

Additionally, two surveys (Simmhan et al., 2005; Bose & Frew, 2005) focus on provenance technologies addressing scientific data and e-science. More specifically, Memento (Van de Sompel et al., 2009) a protocol for finding older versions of Web pages, begins to address the dimension of versioning. In general, Web archiving will be an important infrastructure for addressing these provenance requirements (Masanès, 2006). Other Web models and protocols, such as Dublin Core (Baker, 2005), OAI-ORE (Lynch et al., 2007), and Atom (Gregorio et al., 2007) provide some primitives for describing provenance, which help address the publication dimension. Additionally, OAI-ORE can help to address the problem of collections or grouping of content.

While the Web is an open environment that is different from classic archives, there is much to be gained from the knowledge of archival theory and practice. In particular, a key gap is how to apply the "principle of provenance" (Bearman & Lytle, 1986) in the open world of the Web. Practically, metadata formats stemming from the archival and preservation area, such as the Encoded Archival Context (Szary, 2006), the Metadata Encoding and Transmission Specification, the Metadata Object Description Scheme, and the Preservation Metadata Implementation Strategy data dictionary provide a basis for representing and describing provenance data (Dappert & Enders, 2008). This work can help address the dimension of publication.

Although a number of prior metadata standards already support some aspects of provenance, even librarians and archivists use only around five percent of the fields available in MARC (Moen & Bernardino, 2003); similarly, most users only populate about five of the fifteen core fields available in Dublin Core (Ward, 2003). However, unlike in the classical archiving or library information system scenario, we anticipate Web-based information systems that could automatically populate many of these fields when data is produced via an automatic process, as is already done for scientific workflow management systems (Moreau, 2010; Simmhan et al., 2005). Nevertheless, the question of how to collect adequate provenance, especially from end-users, is a challenging open problem.

## Concluding Remarks

The availability of enormous volumes of information and data in the digital information age makes it crucial that we develop provenance frameworks to capture, manage and use provenance. Without provenance, information is hard to understand, integrate and trust. Although provenance appears on the surface to be a straightforward issue to address in the development of information systems, we have presented a number of important challenges that arise in representing, publishing, accessing and using provenance, and illustrated them with three scenarios of broad applicability. Both researchers and practitioners must address this vitally important area for the future of the Web and for the design of open information systems.

## Acknowledgements

We would like to acknowledge all the contributions by the members of the W3C Provenance Incubator Group in numerous discussions. In particular, we would like to thank Yogesh Simmhan for providing the image for the News Aggregator scenario; Luc Moreau, Yogesh Simmhan, Paolo Missier for providing detailed comments on this document; Jun Zhao, Deborah McGuinness, Oleksiy Chayka, Jose Manuel Gomez Perez, Harry Halpin, Lalana Kagal, James McCusker, James Myers, Luc Moreau, Christine Runnegar and Satya Sahoo for specifying requirements; and Rocio Aldeco-Perez, Jitin Arora, Chris Bizer, Aleksey Chayka, Kai Eckert, Andre Freitas, Irini Fundulaki, Olaf Hartig, Laura Hollink, M. Scott Marshall, Paolo Missier, Luc Moreau, Jim Myers, Iman Naja, Joshua Phillips, Paulo Pinheiro da Silva, Ronald Reck, Marco Roos, Satya Sahoo, Amit Sheth, Marieke van Erp, Brent Weatherly and Jun Zhao for contributing use cases in which the scenarios presented here were built. We would also like to thank Ivan Herman, Thomas Roessler, and Harry Halpin of W3C for their support of this activity.

## References

Artz, D. & Gil, Y. (2007). A survey of trust in computer science and the semantic web. *Journal of Web Semantics, 5*(2). Elsevier, Amsterdam, The Netherlands.

Baker, T. (2005). A common grammar for diverse vocabularies: The abstract model for Dublin Core. Digital Libraries: Implementing Strategies and Sharing Experiences. *Lecture Notes in Computer Science, 3815/2005*(495). Retrieved from doi:10.1007/11599517_71

Bearman, D. & Lytle, R.H. (1986). The power of the principle of provenance. *Archivaria, 21*. The Association of Canadian Archivist, Ottawa, Canada.

Berman, F. (2008). Got data? A guide to data preservation in the information age. *Communications of the ACM, 51*(12). ACM, New York, NY, USA.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information Systems (IJSWIS), 5*(3). IGI Global, New York, NY, USA.

Bose R. & Frew, J. (2005). Lineage retrieval for scientific data processing: A survey. *ACM Computing Surveys, 37*(1). ACM, New York, NY, USA.

Buneman, P., Cheney, J., Tan, W.C. & Vansummeren, S. (2008). Curated databases. Paper presented at the 27th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. Vancouver, BC.

Cheney, J., Chiticariu L. & Tan, W.C. (2009). Provenance in databases: Why, where and how. *Foundations and Trends in Databases, 1*(4). NOW, Hannover, MA, USA.

Dappert, A. & Enders, M. (2008). Using METS, PREMIS and MODS for archiving eJournals. *D-Lib Magazine, 14*(9/10). Corporation for National Research Initiatives, Reston, VA, USA.

Gregorio, J. & de hOra, B. (2007). *RFC 5023: The Atom Publishing Protocol.* Report of the Internet Engineering Task Force. Retrieved from http://www.ietf.org/rfc/rfc5023.txt

HM Treasury. (2003). The green book: Appraisal and evaluation in central government. Retrieved from http://www.hm-treasury.gov.uk/data_greenbook_index.htm

Lynch, C., Parastatidis, S., Jacobs, N., Van de Sompel, H. & Lagoze, C. (2007). The OAI-ORE effort: progress, challenges, synergies. Paper presented at the 7th ACM/IEEE-CS Joint Conference on Digital libraries. Retrieved from http://dx.doi.org/10.1145/1255175.1255190

Masanès J. (2006). *Web archiving.* Springer-Verlag New York Inc., Secaucus, New Jersey, USA.

Moen, W.E. & Benardino, P. (2003). Assessing metadata utilization: an analysis of MARC content designation use. Paper presented at the 2003 International Conference on Dublin Core and Metadata Applications: Supporting Communities of Discourse and Practice. Seattle, Washington, United States.

Moreau, L. (2010) The foundations for provenance on the Web. *Foundations and Trends in Web Science 2*(2-3). NOW, Hannover, MA, USA.

Simmhan, Y.L., Plale, B. & Gannon, D. (2005). A survey of data provenance in e-science. *ACM SIGMOD Record, 34*(3). ACM, New York, NY, USA.

Szary, R.V. (2006). Encoded archival context (EAC) and archival description: Rationale and background. *Journal of Archival Organization, 3*(2). Taylor & Francis, Inc., Philadelphia, PA, USA.

Van de Sompel, H., Nelson, M.L., Sanderson, R., Balakireva, L., Ainsworth, S. & Shankar, H. (2009). Memento: Time travel for the web. Retrieved from http://arxiv.org/abs/0911.1112

Ward, J. (2003). A quantitative analysis of unqualified Dublin Core metadata element set usage within data providers registered with the Open Archives Initiative. Paper presented at the 2003 ACM/IEEE Joint Conference on Digital Libraries, Houston, TX, USA.