# Provenance for Web 2.0 Data

Meghyn Bienvenu, CNRS and Université Paris-Sud
Daniel Deutch, Ben Gurion University of the Negev
Fabian M. Suchanek, Max-Planck-Institute for Informatics

**Abstract.** In this paper, we look at Web data that comes from multiple sources, as in the Web 2.0. We argue that Web data is more than just its content. Rather, a piece of Web data carries along different facets, such the *transformations* that data underwent, the different *perspectives* that users have on the content, and the *context* in which a statement is made. We put forward the idea that *provenance*, i.e. the tracing of where data comes from, can help us model these phenomena. We study how far existing approaches address the issue of provenance for Web data, and identify gaps and open problems.

## 1 Introduction

With the arrival of the Web 2.0, virtually everyone can publish and disseminate data on the Web. This phenomenon contributes to the richness and diversity of content on the Web. At the same time, it gives Web data dimensions that go beyond its pure content: Every piece of information carries a history of where it was first produced, by whom it was first produced, and in which context it was first stated. In this article, we shed light on these attributes of Web data, and we make first steps towards a formal model for these phenomena. We start by exemplifying some of the key properties of Web data.

**Data transformation.** Consider a social network such as Facebook or Twitter: users can publish their own opinions and knowledge, but they can also refer to data posted by other users. Such references typically propagate further, through friends of friends. This phenomenon applies also to other medias such as blogs, emails, and collaborative resources like Wikipedia. Transformations on the Web may be diverse and complex, and "copying" or referring to an existing piece of data is only one example. For instance, there are many services that *aggregate* different feeds into a single piece of data, e.g. Facebook notifications on the number of friends attending an event, or the number of friends "liking" a particular fact. Any given piece of information can carry along an entire history of such transformations.

**User Perspectives.** People may have different views and opinions on the same thing (e.g. on politicians, or the quality of restaurants or hotels). Statements made by one content provider do not necessarily coincide with statements made by another provider. A consumer of data might prefer some providers to others. Thus, the origin or *perspective* of the data is an essential meta-property of Web data.

**Data Context.** The correctness of data (and the trust of users in it) may also depend on the *context* in which this data was published. As a simple example, the statement "Sarkozy is the president of France" is interpreted as true if published in 2008, but is incorrect if published in late 2012. In this simple example, the context of a fact is a timestamp; however, in general, it may be any metadata, such as authorship or location, affecting the trust in facts. We observe that the context metadata is typically not an inherent part of the fact itself. However, it is an essential attribute of any piece of information.

The landscape becomes more intricate when transformation, perspectives, and context are combined: authors may cite other content, and the context of the data should capture both the original and the citing author. This requires the management of context for *propagated* content.

**A Proposed Tool: Provenance.** We propose to model the different aspects of pieces of information by *provenance*. The provenance of a piece of data is a record of meta-information, which is attached to the piece of data, and which indicates where the piece of data comes from. In particular, this meta-information can record the context in which the data was created. In the setting of relational and semi-structured databases, provenance [1–5] was shown to be an extremely useful technique for managing both the original context of data and the ways in which it was manipulated and transformed. We believe that provenance can be used with similar success for Web data, but one of the main challenges here lies in capturing context, perspectives and transformations in a mathematical provenance model.

**Applications of Provenance.** A formal framework for the provenance for Web data can have far-reaching applications. First, it can establish the **authorship** of a certain piece of information. This is useful, e.g., for the protection of intellectual property rights. Provenance can also help guarantee the **privacy** of information (as shown in the context of relational databases by e.g. [1]), e.g in social networks. Finally, provenance is essential for determining the **trust** in a given piece of data.

**Desiderata.** To realize the great potential of provenance for Web data, two main challenges should be addressed. First, one must devise an expressive provenance **model**. This model should account for different types of Web data (Web 2.0, social networks, weblogs, etc.), different kinds of provenance (privacy, location, time, etc.) under different kinds of transformations (references, aggregation, negation, etc.). Principled and generic models for provenance have proven to be highly effective in the context of relational and semi-structured databases; we believe they could play a similarly central role for Web data.

Second, the provenance model should be accompanied by a **reasoning** mechanism. Inspired by [6], we propose that such a mechanism should handle (at least) the following two archetypical questions: (1) *Given a provenance annotation, which statements hold for it?* This would allow queries such as "What happened in 2012?" or "What is the set of statements that both Alice and Bob can access?". (2) *Given a statement, what are the provenance annotations for which the statement holds?* This would allow questions such as "How did this statement come about?", or "Who has sufficient credentials to see this statement?". Note that provenance annotations could be arbitrarily complex: the answer to the question could be "Either Alice or David and Charles together", or "Everybody who is in Unix group X and is friends or relatives with Alice".

## 2 Modeling Provenance and Context

**Provenance in Databases.** Several different provenance management techniques have been introduced in e.g. [3–5]. A general framework for provenance management was proposed in [1]. It uses the mathematical structure of *semirings* to capture provenance of various kinds (including those mentioned above [7]). The model applies to a wide range of applications, in a manner that corresponds exactly to the operations of the positive relational algebra. Extensions of the framework to XML query languages [8] and Datalog [1] have also been defined. Recent work [9] has shown that if the set of available data transformations includes an *aggregate* construct (a common type of Web data transformation), then the semiring framework no longer suffices to capture all possible transformations; an alternative construction based on semi-modules was proposed.

**Provenance in the Semantic Web.** The Semantic Web captures provenance by Named Graphs [10]. Named Graphs equip every triple statement on the Semantic Web with a $4^{th}$ component, the graph identifier, which can be used for grouping triples into sets. Newer versions of the query language SPARQL allow targeting triples of specific sets or asking for sets with specific properties. The Named Graph model is very useful for the Semantic Web, but is, by itself, far from being a universally applicable provenance model for Web data. It provides no sophisticated reasoning capabilities, nor any support for transformations beyond simple inclusion. The work of [11, 12] devises provenance for SPARQL queries on linked Web data, and [13] studies algebras of provenance annotations for RDFS. However, a general model of provenance would need to support also the data transformations that happen in social networks or in collaborative platforms such as Wikipedia. The work on watermarking ontologies [14, 15] serves to prove the provenance of Semantic Web data. However, it does not provide a model of provenance in general, let alone means to reason on it.

**Provenance on the Web.** A large amount of data on the Web already comes naturally with provenance information. The content of every Web page is trivially associated with its URL. Information quoted from or taken from other pages often comes with an indication of the source. Social networks know the concept of authorship, which translates directly into provenance in our setting. Information that has been extracted automatically from Web pages provides another source of provenance data. Many extraction systems [16–19] note from which source a piece of information was extracted. In a similar spirit, automatically generated ontologies [20, 21] can often trace the source of every one of their statements. The YAGO ontology [20], for example, has systematically attached provenance meta-data to its triples. YAGO stores the source, the confidence of extraction, and the extraction technique with every single one of its facts – totaling 80 million for the entire ontology. Recently, YAGO's facts have been annotated with a temporal and a geo-spatial component [22], indicating when and where an event took place. This yields hundreds of thousands of statements with attached meta information, making it an ontology that is anchored in time and space. This work can serve as a use-case for the model of Web data provenance.

**Context and Viewpoints in Artificial Intelligence.** J. McCarthy was among the first to highlight the need for a formal treatment of context [23]. He proposed [24] to annotate formulas with the context in which they hold, i.e., to use formulas of the form $ist(c, \varphi)$, which mean that $\varphi$ is true in the context $c$. He argued that contexts should be first-class citizens in the logic, enabling statements about the properties of particular contexts and the relationships between them, e.g. "If someone is the president of a country, then that person is president in all geographic subcontexts". McCarthy's ideas inspired several concrete logics of context [25–28]. An alternative framework, called multi-context systems [29, 30], treats contexts as local theories which can be interrelated by means of bridge rules, which specify how information can be transferred between contexts.

Epistemic logics [31, 32] have long been studied as a means for representing and reasoning about the viewpoints of different agents. Such logics augment a standard logic, most commonly propositional logic, with a set of *epistemic operators* which can be used to make complex statements about the viewpoints of a group of agents, such as: "Alex does not know that Sue and Bob are dating, but Mary believes that Alex knows that they are dating". Various extensions [33] have been proposed to capture the dynamics of knowledge and belief (e.g. that once Bob tells Alex that he dates Sue, Alex now knows this).

Context logics and epistemic logics were not designed to handle large amounts of *data* and do not offer database-style querying capabilities. Moreover, in the case of multi-agent epistemic logics, the basic reasoning tasks are PSPACE-hard [34], making them ill-suited for querying vast quantities of Web data.

**Challenges.** While the different provenance models described above have proven useful in their respective fields, none is able to address all the desiderata for Web data provenance. If we take the models used in databases as a starting point, we notice that the transformations of provenance are mirrored in the operations on the data (e.g. semiring operations correspond exactly to the operations of the positive relational algebra). In order to extend this idea to the realm of Web data, the following challenges must be addressed: (1) Identify a set of Web data transformations that is expressive enough to capture common features of Web data; (2) Design a mathematical model that captures these transformations; and (3) Design an automated mechanism that identifies real-life transformations, and outputs an instance of the model. The envisioned model should allow complex inference on contexts and viewpoints, as is done in AI, and it should be semantic, large-scale, and distributed - issues addressed in (Semantic) Web research. We believe that a holistic approach bridging these different research areas can yield a model for Web data provenance that is both expressive and tractable.

## 3 Reasoning about context

**A Simple Model.** To illustrate possible reasoning tasks on provenance, we present a simple model for contexts and propagation. We define a *context-*

*annotated database* as a pair $(D, A)$, where $D$ is a database, and $A$ is a mapping from the facts of $D$ to a positive Boolean expression over a fixed set C of context variables. The semantics is given in terms of valuations of the context variables. Given a valuation $V$ of C, we denote by $V(D, A)$ the (standard) database which contains exactly those facts $f \in D$ such that the Boolean expression $A(f)$ evaluates to true under $V$. In addition, we allow a *background theory.* This is a propositional formula (including negation) over C that captures basic relationships between contexts. Finally, the user may be interested only in certain contexts, which may be the conjunction or disjunction of atomic contexts from C (e.g., all facts that hold in 1990 or 2000). We thus define the notion of a *viewer context*, as a positive propositional theory over C.

**Data Transformations.** We add a simple model for data transformations, captured by positive relational algebra transformations of the data; these correspond to positive Boolean algebra on annotations. For example, if the same rumor is sent to Alice by two friends Bob and Carol, then it is associated with a Boolean expression `Bob` ∨ `Carol` (where `Bob` and `Carol` are context variables standing for trust in Bob and Carol respectively), expressing that it is enough for Alice to trust one of them in order to trust the fact.

**Reasoning Tasks.** In Section 1, we introduced two general reasoning questions for provenance In our toy model, the decision problems corresponding to these questions can be formalized as follows:

**Definition 1 (POSS Problem).** *Given an annotated database $(D, A)$, a propositional formula $F$ (comprising the background theory and viewer context), and a Boolean positive relational algebra query $Q$, decide whether there exists a valuation $V$ which satisfies $F$ and is such that $Q$ holds in $V(D, A)$.*

**Theorem 1.** `POSS` *is PTIME if $F$ is positive, and NP-complete in general.*

**Definition 2 (CERT Problem).** *Given an annotated database $(D, A)$, a propositional formula $F$ (comprising the background theory and viewer context), and a Boolean positive relational algebra query $Q$, decide whether it is the case that $Q$ holds in $V(D, A)$ for every valuation $V$ which satisfies $F$.*

**Theorem 2.** `CERT` *is PTIME if $F$ is Horn, coNP-complete in general.*

**Challenges.** The above model is merely a toy model designed to illustrate the concept of reasoning on provenance. When moving to more intricate data transformations, notably aggregates, Boolean formulas may not be enough to represent the context [9]. Other transformations (like re-tweeting) might require nesting of annotations, in the spirit of nested modalities from context and epistemic logics, and of provenance for the nested relational calculus [35]. Also, from a usability point-of-view, it may prove more natural to express the annotations and background theory using (a suitable fragment of) first-order logic, rather than propositional logic. Moving to a richer logical language will likely affect the complexity of reasoning, and it might also open up more reasoning tasks beyond the two that we considered. Finally, there are important practical challenges in the implementation of a reasoning engine for provenance; optimizations will be crucial in order to ensure scalability.

## 4 Conclusions

In this paper, we have emphasized the need for a model of provenance for Web data and for reasoning capabilities. We have given an overview of different approaches in the areas of the Semantic Web, in the Database domain, and in Artificial Intelligence. However, we have come to the conclusion that none of the existing approaches addresses the problem of provenance for Web data in its entirety. Thus, we believe that the exploration of the issues of provenance, data transformation, and data perspectives will be a fertile ground for future research.

## References

1. Green, T., Karvounarakis, G., Tannen, V.: Provenance semirings. In: PODS. (2007)
2. Cui, Y., Widom, J., Wiener, J.: Tracing the lineage of view data in a warehousing environment. ACM Transactions on Database Systems **25**(2) (2000)
3. Buneman, P., Cheney, J., Vansummeren, S.: On the expressiveness of implicit provenance in query and update languages. ACM Trans. Database Syst. **33**(4) (2008)
4. Buneman, P., Khanna, S., Tan, W.: Why and where: A characterization of data provenance. In: ICDT. (2001)
5. Benjelloun, O., Sarma, A., Halevy, A., Theobald, M., Widom, J.: Databases with uncertainty and lineage. VLDB J. **17** (2008) 243–264
6. Abiteboul, S., Duschka, O.M.: Complexity of answering queries using materialized views. In Mendelzon, A.O., Paredaens, J., eds.: PODS. (1998) 254–263
7. Green, T.: Containment of conjunctive queries on annotated relations. In: ICDT. (2009)
8. Foster, J., Green, T., Tannen, V.: Annotated XML: queries and provenance. In: PODS. (2008)
9. Amsterdamer, Y., Deutch, D., Tannen, V.: Provenance for aggregate queries. In: PODS. (2011)
10. Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named graphs, provenance and trust. In: WWW '05: Proceedings of the 14th international conference on World Wide Web, New York, NY, USA, ACM (2005) 613–622
11. Theoharis, Y., Fundulaki, I., Karvounarakis, G., Christophides, V.: On provenance of queries on semantic web data. IEEE Internet Computing **15**(1) (2011) 31–39
12. Cheney, J., Chong, S., Foster, N., Seltzer, M.I., Vansummeren, S.: Provenance: a future history. In: Proc. of OOPSLA. (2009)
13. Theoharis, Y., Fundulaki, I., Karvounarakis, G., Christophides, V.: On provenance of queries on semantic web data. IEEE Internet Computing **99**(PrePrints) (2010)
14. Suchanek, F.M., Gross-Amblard, D.: Adding fake facts to ontologies. In: Demo at the International World Wide Web Conference, ACM (2010)
15. Suchanek, F.M., Gross-Amblard, D., Abiteboul, S.: Watermarking for ontologies. In: ISWC. (2011)
16. Etzioni, O., Cafarella, M., Downey, D., Kok, S., Popescu, A.M., Shaked, T., Soderland, S., Weld, D.S., Yates, A.: Web-scale information extraction in knowitall: (preliminary results). In: World Wide Web Conference. (2004)

17. Suchanek, F.M., Sozio, M., Weikum, G.: SOFIE: A Self-Organizing Framework for Information Extraction. In: International World Wide Web conference (WWW 2009), New York, NY, USA, ACM Press (2009)
18. Nakashole, N., Theobald, M., Weikum, G.: Scalable knowledge harvesting with high precision and high recall. In: WSDM. (2011)
19. Carlson, A., Betteridge, J., Wang, R.C., Jr., E.R.H., Mitchell, T.M.: Coupled semi-supervised learning for information extraction. In: Proceedings of the Third ACM International Conference on Web Search and Data Mining (WSDM 2010). (2010)
20. Suchanek, F.M., Kasneci, G., Weikum, G.: YAGO: A core of semantic knowledge - unifying WordNet and Wikipedia. In Williamson, C.L., Zurko, M.E., Patel-Schneider, Peter F. Shenoy, P.J., eds.: World Wide Web Conference, Banff, Canada, ACM (2007) 697–706
21. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.G.: DBpedia: A Nucleus for a Web of Open Data. In: International Semantic Web Conference. (2007) 722–735
22. Hoffart, J., Suchanek, F.M., Berberich, K., Weikum, G.: Yago2: a spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence Journal (2012)
23. McCarthy, J.: Generality in artificial intelligence. Communications of the ACM **30**(12) (1987) 1029–1035
24. McCarthy, J.: Notes on formalizing context. In: IJCAI. (1993) 555–562
25. Buvac, S., Mason, I.A.: Propositional logic of context. In: Proc. of AAAI. (1993) 412–419
26. Buvac, S.: Quantificational logic of context. In: Proc. of AAAI. (1996) 600–606
27. Nossum, R.: A decidable multi-modal logic of context. J. Applied Logic **1**(1-2) (2003) 119–133
28. Klarman, S., Gutiérrez-Basulto, V.: Two-dimensional description logics for context-based semantic interoperability. In: AAAI. (2011)
29. Giunchiglia, F., Serafini, L.: Multilanguage hierarchical logics, or: How we can do without modal logics. Artificial Intelligence **65**(1) (1994) 29 – 70
30. Serafini, L., Bouquet, P.: Comparing formal theories of context in ai. Artificial Intelligence **155**(1-2) (2004) 41–67
31. Hintikka, J.: Knowledge and Belief. Cornell University Press (1962)
32. Fagin, R., Halpern, J.Y., Moses, Y., Vardi, M.Y.: Reasoning About Knowledge. MIT Press (1995)
33. van Ditmarsch, H., van der Hoek, W., Kooi, B.: Dynamic Epistemic Logic. Springer (2007)
34. Halpern, J.Y., Moses, Y.: A guide to completeness and complexity for modal logics of knowledge and belief. Artif. Intell. **54**(2) (1992) 319–379
35. Foster, J.N., Green, T.J., Tannen, V.: Annotated xml: queries and provenance. In: PODS. (2008) 271–280