Language-integrated Provenance in Links

Stefan Fehrenbach    James Cheney
University of Edinburgh
stefan.fehrenbach@ed.ac.uk    jcheney@inf.ed.ac.uk

Abstract

Today’s programming languages provide no support for data provenance. In a world that increasingly relies on data, we need provenance to judge the reliability of data and therefore should aim for making it easily accessible to programmers. We report our work in progress on an extension to the Links programming language that builds on its support for language-integrated query to support where-provenance queries through query rewriting and a type system extension that distinguishes provenance metadata from other data. Our approach aims to work solely within the language implementation and thus require no changes to the database system. The type system together with automatic propagation of provenance metadata will prevent programmers from accidentally changing provenance, losing it, or misattributing it to other data.

1. Introduction

Provenance information, knowing where data came from, is crucial in a variety of situations. Over the last few years, researchers extended databases to store, query, and propagate provenance information. Such techniques are strongly motivated by applications to data integrity and authenticity in open, collaborative settings, such as Web databases or data integration systems where data are combined from numerous sources of variable quality.

At the same time, advances in language-integrated query bridge the gap between relational databases and programming languages [9, 16]. However, so far these avenues of research seem to take little notice of each other. In particular, there has been little investigation of how to write programs that interact with a database that provides support for provenance. Most applications where provenance is needed do not just involve a database in isolation; for example, a typical web application usually involves at least three separate layers, each controlled by a different language such as server-side middleware (e.g. Java, Python, PHP), database (SQL), and Web client (JavaScript). We think the interaction of programming languages with provenance can be greatly improved by leveraging ideas from language-integrated query and Web programming.

We extend the research programming language Links [9] with explicit support for provenance. Links was originally designed to support writing three tier web applications in a single language. Links programs can run on the client, the server, and query the database using list comprehensions that translate to SQL; furthermore, its type system statically checks that the communication between these layers is well-behaved. Links provides type-safe language-integrated querying, as described elsewhere [8, 15]. In particular, the type system of Links checks statically that embedded query expressions will generate a single SQL query.

To illustrate, imagine the following scenario: Consider Pear Computers, a major international company specializing in high-end music players, smartphones, and smartwatches. We are developing a website on behalf of Pear Computers that features customers’ comments to promote products. Figure 1 illustrates this scenario using Links. We have a table top_comments that aggregates comments on all Pear Computers products from a variety of sources and somehow keeps track of where every comment came from. The table declaration tells Links about the existence of the table, and the names and types of its columns. On the product page for Pear’s new and highly anticipated smartwatch, the pWatch, we only want to show comments that relate to watches, not phones or other products. The watch_comment function identifies comments as relating to watches by a crude heuristic: they originate from a table called watch or contain the word pWatch. We render a single quote with the render_quote function which makes use of Links’ literal XML support. The quotes_list function ties it all together. It queries the database with a query block that selects the text of every comment in top_comments that is a watch comment. The query is guaranteed to generate a single SQL query at run time: it calls a function watch_comment that only performs operations that are allowed on the database. Finally, it returns the XML representation of a list of all the watch comments, which can be embedded in an HTML page by other parts of the program.

Now suppose that we would like to be able to remove comments that do not comply with the Pear Computers company policy of only ever saying how innovative and awesome their products are. In the admin panel, we would like to render comments with a delete button next to them. Obviously, we have to track where the comments came from to be able to delete them. Also, keep in mind that we are dealing with a legacy system in our example: The top_comments table is automatically generated, so it would not help us to delete rows from there. Instead, we have to delete the original data. Fortunately, someone thought ahead and included the necessary information in the originate columns. However, to add the delete button in Links we would now need to change several parts of the program to propagate the origin information to the place where it is needed.

This seems like a perfect fit for where-provenance [2–4], in which each data value in the result of a query carries an annotation consisting of either a reference to the source of the data in the input (e.g. a triple containing the table name, field name and row id), or a placeholder called “bottom” (⊥) indicating that the value was

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page.

TaPP 2015, July 8–9, 2015, Edinburgh, Scotland.
Copyright remains with the owner/author(s).
The keyword `prov` gives access to provenance. It returns a record with three fields: relation, column, and row that correspond to the three components of a where-provenance triple. As before, we consider comments to be about a watch, if they originate from a relation called `watch` (now accessed using `(prov c).relation`) or the text itself (accessed via the keyword `data`) contains the word `pWatch`. We add a new function named `delete_quote` which uses the text’s provenance to delete it from the original table.² This function is called when admins click the delete button that is emitted in the database system, we get provenance support for any unmodified relational database management system for free.

This paper describes work in progress. In the next section, we describe our technical strategy and summarize experience with a preliminary implementation. Section 3 outlines our revised design based on this experience. Section 4 describes related work, Section 5 details limitations and our future plans, and Section 6 concludes.

## 2. Technical approach and preliminary results

Query expressions in Links are based on the nested relational calculus [5], a core query language that provides collection types and comprehensions:

\[
\begin{align*}
    e &::= e \mid (e_1, e_2) \mid e.i \mid e_1 + e_2 \mid e_1 = e_2 \mid \cdots \\
    &\quad | \text{if } e \text{ then } e_1 \text{ else } e_2 \\
    &\quad | \emptyset \mid e_1 \cup e_2 \mid e \mid \{x \leftarrow e\} \mid \text{return } e' \\
    \tau &::= b \in \text{int, bool, \ldots } | t_1 \times t_2 \mid \{\} \\
\end{align*}
\]

Links queries use a similar syntax (extended with records) and uses a normalization algorithm (explained by Cooper [8] and by Lindley and Cheney [15]) to turn such query expressions into SQL, provided the queries return flat records of values of base types. (Recent work on *query shredding* [6] shows how to lift this restriction, but hasn’t yet been incorporated into the main version of Links.) Buneman et al. [3] proposed a translation that maps an ordinary nested relational query to one that propagates where-provenance information on all parts of the source data to the output. We present a simplified version of this translation: given a query expression `e`, we define `P(e)` to be a query that propagates annotations on data values of base types

²We use `table_from_name` to get the actual table from its name.
Figure 3. From PLinks to SQL.

(integers, strings etc.) only, as follows:

\[
\begin{align*}
P(x) &= x \\
P(c) &= \langle c, \bot \rangle \\
P(e_1 \text{ op } e_2) &= \langle P(e_1), 1 \text{ op } P(e_2), 1, \bot \rangle \\
P(e \cdot i) &= P(e) \cdot i \\
P((e_1, e_2)) &= \langle P(e_1), P(e_2) \rangle \\
P(\text{if } e \text{ then } e_1 \text{ else } e_2) &= \text{if } P(e) \text{.1 then } P(e_1) \text{ else } P(e_2) \\
P(\emptyset) &= \emptyset \\
P(e_1 \cup e_2) &= P(e_1) \cup P(e_2) \\
P(\{e\}) &= \{P(e)\} \\
\end{align*}
\]

The translation changes the types of the expressions as follows:

\[
T(b) = b \times \text{tag} \quad T(\tau_1 \times \tau_2) = T(\tau_1) \times T(\tau_2) \quad T(\{\tau\}) = \{T(\tau)\}
\]

where \(\text{tag}\) stands for the type of the provenance information, e.g. triples of relation name, field name and row id. Given a query expression \(e : \tau\) with free variables \(x_1 : \tau_1, \ldots, x_n : \tau_n\), its translation \(\epsilon' = P(e)\) has type \(T(\tau)\) assuming the free variables’ types are adjusted to \(x_1 : T(\tau_1), \ldots, x_n : T(\tau_n)\).

We have used this translation as the basis of an initial attempt to extend Links with provenance support. We modified the Links query normalizer to perform the above translation on queries. However, this approach seems to have some limitations that make it difficult to combine provenance-aware queries with Links programs. For example, since the translation changes the types of both the inputs (e.g. table references) and results, using the translated query requires fixing some convention for naming the additional provenance fields of the tables. These must be added somehow to the database, which imposes a space and maintenance overhead; moreover, the translation requires annotations on all source data, and propagates these annotations through the query eagerly, even if the user never requests them. Moreover, Buneman et al.’s approach relies on a special “bottom” value \(\bot\) to annotate parts of the result that were not copied from the input but were instead created by the query. Finally, Buneman et al.’s translation treats the provenance information as pure metadata: it cannot be inspected by the query, nor can it be manipulated or “forged”. However, once the provenance leaves the query and is processed by the Links program, it becomes plain data: the programmer can discard it, change it or fabricate it. Both of these choices are questionable: it may be advantageous to be able to inspect the provenance during a query, and it may also be advantageous to be able to certify the integrity of the provenance information (i.e. unforgeability) outside of the query.

These observations motivate a design that takes the eventual use of the provenance information elsewhere in Links programs into account, which will be the focus of the next section.

3. Design of PLinks

We describe PLinks, an extension of Links, that turns the programming language into a provenance system in its own right. That is, not only will it allow programmers to query provenance from an external database with provenance support, but it will rewrite queries to calculate their own provenance before they even reach the database.

This is work in progress and has not been implemented yet. Figure 3 shows the compilation and execution model. We type check PLinks programs using the type system extension described in Section 3.1. We translate type correct programs to standard Links using the source to source translation described in Section 3.2. From there, Links uses its standard query normalization strategy described by Cooper and by Lindley and Cheney [8, 15], to generate SQL queries which can then be executed on any relational database.

The new language features of PLinks compared to Links are as follows. PLinks adds a new type constructor Prov \(\alpha\), where \(\alpha\) is a base type, to represent values that carry where-provenance. To access provenance, we add the keyword prov. Given a value of type Prov \(\alpha\), it returns a record representing the value’s provenance. Such a provenance record has three fields: relation, column, and row that reflect the first, second, and third component of a where-provenance triple, respectively. The related keyword data drops provenance, returning just the data. We used most of these in the example in Figure 2. Values of type Prov \(\alpha\) cannot be constructed directly by the programmer. Instead, they are automatically generated by PLinks (based on the prov clauses associated with table definitions). The programmer only specifies how the provenance for a column is derived. This is more flexible than in our initial attempt but possibly not quite enough yet. See Section 5 for a discussion of limitations.

3.1 Type system

Links is a statically typed language. It has base types Int, Bool, and String that correspond to their equivalents in databases. Records have row types that describe the labels and types of the record’s fields. Row polymorphism allows the same function to operate on different types of inputs, by the presence of label fields. Effect types specify where function can be executed: on the client, the server, or in database queries. The most important aspect of Links’ type system, for our purposes, is the guarantee that query blocks get translated to a single SQL query if they use only database-executable functions and return a flat list of records with properties of base type. Lindley and Cheney [15] describe Links’ type system in detail.

PLinks extends Links’ type system to account for where-provenance. Where-provenance is attached to “cells” in a relation and describes where the data was copied from. Provenance information has meaning only in the context of the data it belongs to. The type system should capture the special nature of provenance metadata. This allows programmers to rely on the accuracy of provenance and prevents mistakes. There are four things in particular, that we expect from a type system that properly deals with provenance: First, provenance is attached to data and automatically propagated with the data it belongs to. Second, provenance is immutable, thus the type system should prevent accidental modification. Third, changes to the data would invalidate provenance. Data in Links is immutable, so this problem does not directly manifest, but it may require caution when data in the database is updated. Fourth, it is not possible for a programmer to forge provenance. Provenance is always and only automatically derived from the database by PLinks.

Our type system design realizes these goals as follows. The type constructor Prov indicates data paired up with its provenance. Values of type Prov \(\alpha\), where \(\alpha\) is a base type, carry provenance metadata. The restriction of the type argument \(\alpha\) to base types reflects that where-provenance is attached to “cells”, not whole database rows. This does not apply to other forms of provenance (e.g. why-provenance [4] based on row annotations), so we will need to revisit this in future work.

Figure 4 shows simplified versions of the additional type rules we need for PLinks, compared to Lindley and Cheney’s description of Links’ type system [15]. The PROV rule allows us to read a value’s provenance using the prov keyword. If \(M\) is an expression with type Prov \(\alpha\) we can read its provenance. Provenance is returned as a record (abbreviated PR) with fields relation, column, and row. For the time being we represent relation and column by their names and
The row by an assumed numerical primary key. The `data` keyword projects a value with provenance to just the data, thus the `DATA` rule says that the data-part has type `o` if the term has type `Prov o`.

The `TABLE` rule introduces `Prov` types. In Links, the programmer declares a `table` `t` with some columns `li` that contain values of base types `oi`. In `PLinks`, the additional `prov` clause allows programmers to declare a function `f_p` for each of a subset `lp` of the columns `li`. This function will be used to compute the provenance of a value of column `p`. It takes a database row as its input and produces a provenance record (`PR`) and needs to be database-executable. The type of a `table` declaration is list of records with the declared fields.

Fields `li` that appear in the `prov` clause have type `Prov `oi`. Fields `li` without provenance have just type `oi`.

Note that all operations that would introduce “bottom” provenance in Buneman et al.’s presentation of where-provenance are ill-typed in `PLinks`. We have decided against adding a “bottom” constructor and corresponding type rule for now. It would weaken the guarantees the provenance type gives us and it seems like most, if not all, cases where it is useful can be covered by wrapping values in an explicit representation of alternatives, like Haskell’s Either.

### 3.2 Translation

We intend to implement `PLinks`, as indicated in Figure 3, by translating `PLinks` programs into Links programs. Thus, we need to express the new keywords and types of `PLinks` in terms of Links features.

As an example of the translation, we will translate the `query` block in Figure 2. It refers to the `top_comments` table, filters out tuples based on their provenance using the `watch_comment` function, and finally returns the contents of the text column together with its provenance. We translate this piece of `PLinks` code into the Links code shown in Figure 5.

Types of the form `Prov `o` are special during type checking. After that, we replace them by record types with a data and `prov` field. The data field has type `o` and will contain the actual data. The data field has the usual provenance record type and will contain provenance metadata. In the example, we see the translation of `Prov `o` types in the signature of `watch_comment`. Where there used to be `Prov String` in Figure 2, there is a record type in Figure 5.

In the body of `watch_comment` we see how to translate the `PLinks` keywords `prov` and `data`. Corresponding to the translation of the `Prov `o` type, they are simply translated into projections. Note that the less restrictive record types, compared to `Prov `o`, do not enable programmers to circumvent the type system restrictions because we only translate programs that have passed the more restrictive type checker already.

So far, the translation was very straightforward syntactic sugar on top of records. Thus, the meat of the translation has to happen in creating these records. Indeed, we see that where we have just a reference to `top_comments` in Figure 2, we have a whole nested

---

**Figure 4.** Additional typing rules compared to Links [15].

**Figure 5.** Translated query block from Figure 2.

**Figure 6.** SQL query generated for the code from Figures 2 and 5.
4. Related work

To the best of our knowledge, we are the first to propose supporting provenance by translation within a programming language (leveraging language-integrated query support) rather than by altering the database system or extending it with stored procedures [1, 7, 12]. While we have focused on supporting provenance by translation to SQL, our approach also should be able to cope with database-side support for provenance, by generating queries in provenance-aware query languages such as Trio’s TriQL or Karvounarakis et al.’s ProQL [14].

Corcoran et al. [10] developed SELinker, a version of Links extended with label-based security enforcement, including provenance-like label propagation; this was implemented using a theory of type coercions [17]. However, their approach relied on a user-defined type and stored procedures for storing and manipulating sets of labels.

Glavic et al. [13] compiled this list of requirements that a provenance system should implement: (1) support different kinds of provenance, (2) support complex queries, by which they mean a large subset of SQL if not everything, (3) complex queries over provenance data itself, and (4) scale to large databases. Although we have only considered where-provenance for the relatively simple subset of SQL supported by Links to date, we believe our approach can be extended to handle other forms of provenance and to handle richer queries. In contrast to other approaches, we have a type system that restricts operations to those that make sense in the presence of where-provenance. Our approach addresses requirement (3) for where-provenance in a new way. Regarding scalability, we have not carried out a detailed experimental evaluation, and it will be interesting to compare our approach with existing techniques based on changing the database system or extending it with stored procedures.

5. Future work

PLinks as described in Section 3 has not yet been implemented: although we have a preliminary implementation of the where-provenance translation from Buneman et al. [3], we have not yet extended Links with the Prov type constructor, the prov and data keywords, or the prov table modifier as described in Section 3. This is the obvious first step before gathering experience, performance evaluation, and revising the design.

Another area for future extension is database updates. When writing data to a database we can refer to the data’s provenance and thus implement provenance propagation across tables.

In our current design, we make programmers drop provenance explicitly using the data keyword. In the future, we hope to insert it implicitly when needed. This will reduce the amount of code that needs to be adapted when adding provenance to a column.

Links has automatic type inference, but we currently require programmers to annotate functions that deal with provenance. Type inference for provenance might raise the question of polymorphism with respect to provenance.

PLinks uses programmer-defined functions to map a row to a columns provenance. This admits some flexibility in how exactly provenance is stored. It is tempting to think that we can use this to retrieve provenance for a record from another table like this:

```
<table>
<thead>
<tr>
<th>T</th>
<th>r</th>
</tr>
</thead>
<tbody>
<tr>
<td>table</td>
<td>“r” with [id : Int, c : String] prov (c = fun (x) { THE(for (p &lt;-- external_provenance_table) WHERE (p.id == x.id) [(relation = p.relation, column = p.column, row = p.id)])})</td>
</tr>
</tbody>
</table>
```

However, this requires some function THE with type ([r]) -> r for some type record type r that is database-executable. Unfortunately, no function with this type can exist in Links as it is, because it would violate the guarantee that we create a single SQL query from a query block. At its core, this problem seems to be about expressing foreign key relationships. If we were able to tell Links that some columns uniquely identify a row in another table, Links could use that information to just emit a join. How to encode such a restriction in the type system remains is not yet clear to us.

When data of type Prov o leaves a query block, we attach provenance metadata to it, whether it is ultimately needed or not. Where-provenance is only a constant factor in increased memory but other forms of provenance can result in larger overhead. This may turn out to be a real problem. In that case, we might want to consider lazy calculation of provenance. This would still require storing enough provenance to be able to query the full provenance. In particular, this raises the question of how to deal with a database that changes between fetching data and retrieving its provenance.

We also aim to apply the approach described here to other forms of provenance besides where-provenance, such as why-provenance or how-provenance. Prior work by Corcoran et al. [10] and Glavic et al. [12, 13] suggests that this should be possible, we intend to explore extending PLINKs with rewriting-based support for other forms of provenance, possibly in concert with recent work on query shredding [6].

Although Links is well-suited for prototyping provenance support, it is a research language with a small audience. To make our work more accessible and useful, we plan to consider whether it is possible to apply similar ideas to (subsets of) more mainstream languages that support comprehensions or language-integrated query, such as F# or Python.

6. Conclusions

Provenance within (relational and other) databases has been investigated extensively. However, all proposals to date involve significant changes or extensions to the database system and either extend the query language (SQL) or change the query results (or both). We propose instead to support provenance at the programming language level, by translating queries so as to propagate provenance information. This strategy offers a number of possible advantages, including stronger guarantees about the integrity of provenance within the programming language, and the ability to make provenance-aware programs portable across database backends (e.g. either generating plain SQL for mainstream databases, or queries in a provenance-aware query language if the database being used supports it). This paper reports on our preliminary implementation and proposes a refined design based on experience so far; we outline a number of possible directions for future work.

Acknowledgments

This research is supported in part by the by EU FP7 project DIACHRON (grant number 601043) and by a Google Research Award.

References


