Citations, Certificates and Object References*

Christoph Koch
LFCS, University of Edinburgh, UK
koch@dbai.tuwien.ac.at

There are two ways in which objects may refer to other objects. The first is common in databases, and we shall refer to it as an *object reference* or *oref*. For example, if a database contains employees and departments, and we want to represent the fact that Joe works in the Widget department, we place the appropriate oref – a pointer or foreign key value – into the department field of the record associated with Joe. If, at some later time, we ask for the manager of Joe's department, we look for the *most recent* record associated with the Widget department. In all likelihood the database will only store the most recent version.

Now compare this with the situation in which you are writing a paper for publication on the Web. If you refer to another document, you often assume that the document is fixed. However, if that document has several revisions or versions you can – and should – take care to insert a URL to a particular version of the other document, which is likely to be the most recent version of the document at the time you composed your paper. Then, someone following the URL at some later time expects to be taken to an *old version* of the document – the version of the document that you *cited*.

This distinction between orefs and citations is obvious and unremarkable. However, it becomes interesting once we blur the distinction between databases and documents and assume, as is now common in scientific publishing, that a citation may refer some part of a database. This is even more critical to scientific databases: We have situations in which one database may contain citations to another database. Here the distinction between orefs and citations is sometimes forgotten.

If a database is properly archived, then producing a citation to a part of some previous version presents no problems. But in many cases databases are large, heterogeneous and distributed (e.g. the whole Web), such that it is impossible to produce an instantaneous snapshot of the database. In this case, what should we expect to find when we follow a citation?

We propose that when you query a database, it should return not only an answer, but also a *certificate of origin* (or certificate for short) for the answer.

*Joint work with Peter Buneman.

That certificate may be used as a citation and, if you present the certificate to the database at some time in the future, the certificate can be used to provide you with evidence of what data the query looked at but maybe not much more. For example, assume we wanted a *citation* for the name of the manager of Joe's department. The result would be a name together with a certificate for the query. That certificate could be used to find certificates for the other objects involved in the query: the record for Joe and the record for his department. However the certificate may not be used to extract any other past information from the database. It may not tell you about other employees, and it may not tell you about some other attribute values (e.g. Joe's salary). Thus we cannot use certificates for arbitrary historical queries.

We believe that certificates are useful as a basis for recording provenance, and we are currently working to develop a *data model* that includes and is based on the notion of certificates of origin. We want to pursue the following topics:

- The study of the exact relationship between the notion of object identity and certificates of origin.
- The development of a query language that is "conscious" of both orefs and citations.
- The meaning of update in this data model.
- The connection between this language and classical database query languages both with respect to expressiveness and evaluation complexity.
- The study of the overhead required to support this additional machinery.
- Characterizing those historical queries that can be supported by our enhanced data model.
- Given a historical query that cannot be answered on the basis of this framework, can we characterize approximations to the query?
- How does this data model connect with previous approaches to versioning and archiving databases?
- The use of this model in understanding the attachment of annotations to data. Can we identify annotation with (materialized) views based on queries that return a set of certificates in our citation-conscious query language?

I am greatly indebted to Peter Buneman, who provided motivation for the currently ongoing work as well as a wealth of ideas (even if most of them are still in flux). Some also arose from his discussions with David Maier, Sanjeev Khanna and Wang-Chiew Tan and a number of scientists at Edinburgh University. Thanks to all of them.