As the Web makes it increasingly easy to exchange, copy and transform data, the issue of *provenance* – where data had come from and how it was derived – has rapidly become a leading research issue. Provenance has always been important in scholarship, and it is now becoming important to scientists who deal with large and complex data sets; but we do not need scholars or scientists to tell us of its importance. Go to the Web and search for the population of some country. You may well find that it is impossible to find out where a figure came from or how it was derived. derived.

The importance of provenance has been recently been recognized. The International Annotation and Provenance Workshop is devoted largely to *workflow* provenance; there is also a growing body of database research into *data* provenance The first paper in this issue by Wang-Chiew Tan, provides an accessible and comprehensive introduction to these two aspects of provenance. Data provenance has recently emerged as being important to a number of areas of computer science such as annotation, data integration, probabilistic databases, file synchronization, data archiving and program debugging. In May 2007, James Cheney, Nathan Foster and Bertram Ludäscher organized an informal workshop on the Principles of Provenance at the University of Pennsylvania in May 2007. Its purpose was to bring together people working in these areas in order to elicit some underlying principles and models. This issue is based on some of the talks at that workshop.

Provenance and Data Synchronization, by Nathan Foster and Grigoris Karvounarakis, describes two applications involving data replication: the data synchronization system Harmony where provenance tagging is needed to provide independence from order and the data sharing system Orchestra where provenance is important both for trust and for incremental updates. Following this musical demonstration, in *Program slicing and data provenance* James Cheney shows a connection between program slicing where – for debugging purposes – one wants to focus on that part of the program that influenced a specific variable and data provenance where one is interested in that part of the evolution of a database that accounts for the current state of a specific data item.

Update languages, especially the update fragment of SQL, are often dismissed by database researchers because, when measured by their ability to effect transformations of the database, they are less expressive than query languages. In *Recording Provenance for SQL Queries and Updates*, Stijn Vansummeren and James Cheney point out that when provenance is taken into account, update languages can become interesting because they can express more than query languages. This paper describes both an implicit and explicit semantics for provenance for both query and updates in SQL-like languages.

Although the focus of this issue is more on data provenance than workflow provenance, two of our papers argue forcefully that these two topics should not be divorced. The first, *Issues in Building Practical Provenance Systems* by Adriane Chapman and H.V. Jagadish, provides a set of desiderata for provenance recording and describes how simply recording workflow execution may not be adequate for understanding the evolution of data, especially when this has been heavily manipulated. The second, *Provenance in Scientific Workflow Systems* by Susan Davidson *et al* describes an an approach to summarizing workflows and then – through a collection-oriented model of data – analyzing the effect of the individual processes through a stream-based model of processing. In our opinion, connecting workflow and data provenance is the most interesting research challenge in the general field of provenance models.

Finally, in *Copyright and Provenance: Some Practical Problems*, John Ockerbloom shows how, even in the world of fixed digital documents, provenance in is important in determining intellectual property. Traditional copyright law is based on the fact that copying a work (printing it) was a non-trivial task. While traditional law may not have been ideal, it was at least workable. How we adapt it to electronic data sets in which many small pieces of data have been derived and assembled from a large number of other sources is a major challenge; and it is one in which a good model of provenance may be of significant benefit.