

Curated Databases

Peter Buneman
University of Edinburgh
opb@inf.ed.ac.uk

Scientists, notably biologists, are making increasing use of databases to publish both their data and their interpretation of data. These databases are valuable because of the human effort (curation) that goes into their construction and maintenance. They typically consist of a mixture of source data, metadata, annotations, and relevant data that has been extracted from other curated databases.

Current database and data exchange technology does not serve database curation well. In this talk I shall address a number of issues connected with curated databases.

Annotation of existing data now provides a new form of communication between scientists, but conventional database technology provides little support for attaching annotations. I shall show why new models of both data and query languages are needed.

Closely related to annotation is provenance – knowing where your data has come from. This is now a real problem in bioinformatics with literally hundreds of curated databases, most of which contain substantial amounts of data extracted from other curated databases.

Preserving past states of a database – archiving – is also important for verifying the basis of scientific research, yet few published scientific databases do a good job of archiving. Past "editions" of the database get lost. I shall describe a system that allows frequent archiving and efficient retrieval with remarkably little space overhead.

Finally I shall argue that we need a new model of how curated databases are constructed. The idea that such databases are constructed as views of other data through conventional query and update languages is unhelpful, and that formulation of a "copy-and-paste" model of data construction may provide us with better curation tools.