

# The Two Cultures of Digital Curation

Peter Buneman

University of Edinburgh

[opb@inf.ed.ac.uk](mailto:opb@inf.ed.ac.uk)

## Abstract

The United Kingdom has recently created a *Digital Curation Centre* whose purpose is to provide advice on, develop tools for and conduct research on all aspects of Digital Curation. But what is digital curation, and why is it interesting to database researchers? Ask around, and you are likely to find two kinds of people involved in digital curation -at least they call themselves curators and use computers. Moreover, on the face of it, they have almost nothing else in common.

An archivist (A) does the digital equivalent of putting documents in boxes. A is dealing with data generated by other people and is concerned with: appraisal -the selection of what documents to preserve, indexing and classification -the choice of which document to put into which box, and preservation- ensuring that the documents are preserved for posterity. A finds computers extremely useful because all kinds of “digital objects” may be archived, and the internet provides easy access to digital objects.

A scientist (B) does the digital equivalent of publishing a textbook or compendium. B might be a biologist and is publishing data that results from B's experiments or has been collected as a result of B's research into the literature. B's concerns are with organization and integration of data that has been collected from other sources, with the process of annotation of this data and with the publishing and presentation of the data. B finds computers and the internet useful because it is easy to add recent data -one doesn't have to wait for the next paper edition to appear, one can build rather rich representations of the data, and it is easy to publish the data in a form that is accessible to the readers. In fact, B is likely to use some form of database technology.

What do A and B have to do with each other? Quite a lot -and much of it depends on database technology or presents challenges for database research. In building up a catalog of archived data A is already doing something like B, but perhaps with more stable data. B also needs to be concerned with archival issues. Because B has traditionally been more concerned with publishing than preservation, there are now a number of endangered data sets that are potentially important for longitudinal or historical studies.

In this talk I shall describe some of the challenges for database research and the progress that has been made on them: they include

- Data integration. What has database technology delivered? What can we expect it to deliver? And what is wishful thinking?
- Database archiving. Digital objects are typically fixed, but how does A archive B's data, which may change daily or more often? Should A create a new archive after every update?
- Annotation. This is data that is sometimes attached to a database *after* it has been designed and populated. What does B need in order to attach annotations to databases over which B has no control?
- Provenance. This is loosely related to annotation. It is something that A will describe, but is equally important to B. Suppose I point to some element in a database and ask you where it has come from. In B's domain that element has been repeatedly copied from database to database, so even describing the provenance may be a complicated task. Can one do better than provide the transaction logs of all the databases involved?

The first of these is a major topic of concern to a large number of database researchers. I can only provide a brief and opinionated summary. The last three are topics that I and my colleagues have been working on. I shall describe our progress.