Curated Databases

Peter Buneman

University of Edinburgh, UK
opb@inf.ed.ac.uk

Extended Abstract

Most of our research and scholarship now depends on *curated databases*. A curated database is any kind of structured repository such as a traditional database, an ontology or an XML file, that is created and updated with a great deal of human effort. For example, most reference works (dictionaries, encyclopaedias, gazetteers, etc.) that we used to find on the reference shelves of libraries are now curated databases; and because it is now so easy to publish databases on the web, there has been an explosion in the number of new curated databases used in scientific research. Curated databases are of particular importance to digital librarians because the central component of a digital library – its catalogue or metadata – is very likely to be a curated database. The value of curated databases lies in the organisation, the annotation and the quality of the data they contain. Like the paper reference works they have replaced, they usually represent the efforts of a dedicated group of people to produce a definitive description of enterprise or some subject area.

Given their importance to our work it is surprising that so little attention has been given to the general problems of curated databases. How do we archive them? How do we cite them? And because much of the data in one curated database is often extracted from other curated databases, how do we understand the provenance of the data we find in the database and how do we assess its accuracy? Curated databases raise challenging problems not only in computer science but also in intellectual property and the economics of publishing. I shall attempt to describe these.

M. Agosti et al. (Eds.): ECDL 2009, LNCS 5714, p. 2, 2009.