

What the web has done for scientific data – and what it hasn't*

Peter Buneman
School of Informatics, University of Edinburgh

3 Oct, 2005

Abstract

The web, together with database technology, has radically changed the way scientific research is conducted. Scientists now have access to an unprecedented quantity and range of data, and the speed and ease of communication of all forms of scientific data has increased hugely. This change has come at a price. Web and database technology no longer support some of the desirable properties of paper publication, and it has introduced new problems in maintaining the scientific record. This brief paper is an examination of some of these issues.

1 Introduction

Try to imagine the unthinkable: you have lost your internet connection. So you go to the reference shelves of your local library for some information relevant to your work. Perhaps you are interested in demography and want the GDP and population of some country. The chances are that you will find a rather sorry and little-used collection of reference books, most of them relics of the time before the web – only a few years ago – when libraries were the main vehicle for the dissemination of scientific and scholarly information. The reference books have been replaced by databases to which, if they are not open-access, the library has subscribed on your behalf. You would, of course, be much better off using the web. In fact, for scientific data the web has had huge benefits

- it has provided access to much larger and richer data collections;
- the information is much more timely – we do not have to wait for a new edition to be printed to get an up-to-date GDP;
- access to the information is much faster and simpler;
- the information is better presented; and
- as a result, new information sources have been created which both classify scientific data in useful ways and form a vehicle for the communication of scientific opinion.

The impact of the web on the way scientific research is conducted has been enormous. Michael Lesk [?] has argued that it has actually changed the scientific method from “hypothesize, design and run experiment, analyze results” to “hypothesize, look up answer in data base”. Almost all of modern science is now dependent on databases. Biology has led the way in the use of organised data collections to disseminate knowledge (I shall refer to these as databases) but nearly all branches of scientific research are now dependent on web-accessible data resources. Databases are vehicles for publishing data (in fact the databases themselves can be considered publications), and it is often a condition of scientific funding that an appropriate database be generated and made accessible to the scientific community.

*This is the corrected text of an abstract of a talk presented at WAIM2005. It has already been published in the proceedings of that conference and is not for circulation.

All this represents spectacular progress. We should not be upset that the library is no longer the primary vehicle for the dissemination of scholarship. But is it possible that in the rush to place our data “on the web” we are losing some important functions that libraries – whether or not by design – traditionally provided? Consider again your journey to the library.

First, if the reference work you were looking for is not in the library, the chances are that some other library has a copy of it. By having copies of the same work distributed among many libraries, there is some guarantee that the the information is being preserved. Copying has always been the best guarantee of preservation. Now that your data is kept at some centralised database, it is not at all clear that it is in a form appropriate for long-term presentation. Also, the responsibility for keeping the information is now assumed by the organisation that maintains the database rather than being distributed among a number of libraries.

Second, maybe what you were looking for is in a reference book that is updated from time to time. If the library decided not to buy the new edition, at least you could revert to an old edition. Now, if the library drops its subscription to the on-line data, what do you have? This underlines the fact that the economic and intellectual property issues with databases on the web are very different from those that apply to traditional paper-based dissemination of knowledge. However the law that applies to digital data collections is effectively based on the traditional copyright laws.

Third, once you had found the information you were looking for, there was a serviceable method of citing it according to one of a few accepted methods. You could, if needed, localise the information by giving a page number, and the citation could be checked by other people who had access to the cited document. Now it is not at all clear how you cite something you find in a database; and you have no guarantee that it can be checked. Maybe the web site has disappeared, or maybe the database has been updated.

Fourth, the database keeps up-to-date information, but you might want some old information – perhaps the GDP from some past year. The old publications in the library may have this information, but the database does not.

These differences indicate that there are a number of problems with the dissemination of scientific data on the web. Having fast access to up-to-date research material may come at the price of data quality. Arguably, the web is losing the scientific record as fast as it is creating it; and users of web data have little faith in it until they can verify its provenance and authorship.

The rest of this paper is an attempt to show that, in trying to remedy these drawbacks of web data, we are led to some new and challenging problems that involve databases, the web and other areas of computer science.

1.1 Scientific Data

The use of database technology – in a broad sense of the term – to support scientific data is increasing rapidly. However, scientific data puts different demands on database technology. For example, transaction rates are typically lower in the maintenance of most scientific databases. Scale [?] is arguably important, and complexity is surely important. Not only is “schema” complexity high, but the kinds of interactions between query languages and scientific programming require relatively complex algorithms and place new demands on the already complex area of query optimisation [?]. The latter paper deals well with some of the issues of scale. In this note I want to deal with a largely orthogonal set of issues that have arisen in discussion with scientists who are dealing with databases in which the primary issues, at least for the time being, do not concern scale, but involve the manipulation, transfer, publishing, and long-term sustainability of data. Biological data has been the prime mover in some of these issues, but other sciences are catching up.

2 Data transformation and Integration

Data integration, of course, a relatively old topic in database research, which is crucial to curated databases. While low-level tools such as query languages that can talk to a variety of data formats and databases are

now well-developed; declarative techniques for integration and transformation based on schema annotation and manipulation have been slow to come to market; and where progress has been made it is with relatively simple data models[?, ?]. A survey of the *status quo* is well beyond the scope of this paper, but it is worth remarking that while the emergence of XML as universal data exchange format may help in the low-level aspects of data integration through the use of common tools such as XQuery [?], it is not at all clear whether XML, has helped in the higher-level, schema based approach to data integration. The complexities of constraint systems such as XML Schema [?] appear to defy any attempt at schema-based integration. Moreover, it is not clear what serialisation formats – upon which XML Schema is based – have to do with the data models in which people naturally describe data.

A more limited goal for XML research, and one in which progress has been made, is that of *data publishing*. There is again a growing literature on this topic. The idea is that individual organisations will maintain their data using a variety of models and systems but will agree to common formats, probably XML, for the exchange of data. The problem now is to export data efficiently in XML and, possibly, to transform and import the XML into other databases. This not only requires efficient and robust tools for describing and effecting the transformation [?, ?] but also tools for efficiently recomputing differences when the source database is modified – a new form of the view maintenance problem.

3 Data Citation

A common complaint from people who annotate databases based on what is in the printed literature is that citations are not specific enough. For example, in the process of verifying an entry in some biological database, one needs to check that a given article mentions a specific chemical compound. Even if one is given a page number, it can be quite time consuming to pinpoint the reference. The point here is that the more we can localise citations the better. Now consider the issues involved with citing something in a database. There are two important issues.

- How does one cite the database itself and localise the information within the database?
- How does one cope with the fact that the database itself changes? Does a change necessarily invalidate the citation?

As we know, URLs and URIs fail to meet the needs of stable “coarse grained” identifiers, such as identifiers of a database or web site. This has led people interested in long-term preservation to consider a variety of techniques for maintaining digital object identifiers that persist over an extended period. But even when the domain of citation is in our control, for example we want to specify localised citations within a website or database, how do we specify these citations, and what whose responsibility is it to deal with changing data? For relational databases, a solution to the localisation problem is to use keys or system tuple identifiers. For hierarchical data such as XML, a solution is suggested in [?]. However these are partial solutions to the localisation problem. Standards for data citation need to be developed, and dealing with change is a major problem.

4 Annotation

This is a growing area of activity in scientific databases. Some biological databases describe themselves as “annotation databases”, and there are some systems[?] which are designed to display an overlay of annotations on existing databases. Database management systems typically do not provide “room” for *ad hoc* or unanticipated annotation, and only recently has any attempt been made to understand what is required of database systems to provide this functionality [?]

5 Provenance

Also known as “lineage” and “pedigree”[?, ?], this topic has a variety of meanings. Scientific programming often involves complex chains or workflows of computations. If only because one does not want to repeat some expensive analysis, it is important to keep a complete record of all the parameters of a workflow and of its execution. This is sometimes known as “workflow” or “coarse-grained” provenance [?].

In curated databases, as we have already noted, data elements are repeatedly copied from one database to the next. As part of data quality know where a data element has come from, which databases it has passed through, and what actions or agents created and modified it. Even formulating a model in which it is possible to give precise definitions to these is non-trivial. Moreover, since much copying is done by programs external to the databases or by user actions, it is a major challenge to create a system in which provenance is properly recorded. It involves much more than database technology. For example, data is often copied by simple “copy-and-paste” operations from one database to another. To provide proper mechanisms for tracking this data movement will involve not only re-engineering the underlying database systems but also to the operating systems and interactive environments in which the changes are made.

6 Preservation

Keeping the past states of a database is an important part of the scientific record. Most of us have been “burned” by a reference to a web page or on-line publication that has disappeared or has been changed without any acknowledgment of the previous version. This is one area in which we have made some progress [?]. A system has been implemented which records every version of a database in a simple XML file. For a number of scientific databases on which this has been tested, the size of an archive file containing a year’s worth of changes exceeds the size of one version of the database by about 15%. Yet any past version of the database may be retrieved from the archive by a simple linear scan of the archive.

The principle behind this system is that, rather than recording a sequence of versions, one records one database with the changes to each component or object recorded at a fine-grained level. This relies on each component of the database having a canonical location in the database, which is described by a simple path from the root of the database. It is common for scientific data to exhibit such an organisation, and this organisation may be of use in other aspects of curated data such as annotation, where some notion of “co-ordinate” or “key” is needed for the attachment of external annotation. In fact, it relies crucially on a system for fine-grain citation such as that advocated in section ??.

Archiving in this fashion does a little more than “preserve the bits”. For a relational database, it dumps the database into XML making it independent of a specific database management system and intelligible to someone who understands the structure of the data. The subject of digital preservation is more than preserving bits. It is about preserving the *interpretation* of a digital resource. For example, you have a document in the internal format of an early word processor. Should you be concerned about preserving the text, the formatted text, or the “look and feel” of the document as it was to the users of that word processor [?, ?, ?]. Databases may be in a better position because there is arguably only one level of interpretation – the SQL interface for relational databases, or the syntactic representation of the data in XML. But this does not mean that there is nothing to worry about. How do you preserve the schema, and are there other issues of context that need to be maintained for the long-term understanding of the data?

7 Database research is not enough

Integration, annotation, provenance, citation, and archiving are just a few of the new topics that have emerged from the increasing use of curated databases. Some progress can be made by augmenting existing database technology. But fully to deal with provenance and integration, we need a closer integration of databases with programming languages and operating systems. These require better solutions to the impedance mismatch problem. Some progress was made with object-oriented databases, and more recently in programming

languages [?] and web programming [?] which understand typed data and in which file systems are replaced by database systems as a fundamental approach to solving the impedance mismatch problem. It is not clear whether the natural inertia in software development will ever allow such a radical change to take place. In addition, even if all the technical problems are solved, the social, legal and economic problems with web data are enormous.

References

- [BCF⁺02] Michael Benedikt, Chee Yong Chan, Wenfei Fan, Rajeev Rastogi, Shihui Zheng, and Aoying Zhou. DTD-directed publishing with attribute translation grammars. In *VLDB*, 2002.
- [BDF⁺02] Peter Buneman, Susan Davidson, Wenfei Fan, Carmem Hara, and Wang-Chiew Tan. Keys for XML. *Computer Networks*, 39(5):473–487, August 2002.
- [BKT00] Peter Buneman, Sanjeev Khanna, and Wang-Chiew Tan. Data Provenance: Some Basic Issues. In Sanjiv Kapoor and Sanjiva Prasad, editors, *Proceedings of FST TCS 2000: Foundations of Software Technology and Theoretical Computer Science*, pages 87 – 93. Springer, LNCS 1974, Dec 2000.
- [BKTT04] Peter Buneman, Sanjeev Khanna, Keishi Tajima, and Wang-Chiew Tan. Archiving scientific data. *ACM Transactions on Database Systems*, 27(1):2–42, 2004.
- [CLB01] James Cheney, Carl Lagoze, and Peter Botticelli. Toward a theory of information preservation. In *5th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2001)*, Darmstadt, 2001.
- [Com] C_ω. <http://research.microsoft.com/Comega/>.
- [CTG04] L. Chiticariu, W-C. Tan, and G.Vijayvargiya. An annotation management system for relational databases. In *VLDB*, 2004.
- [CW01] Yingwei Cui and Jennifer Widom. Lineage tracing for general data warehouse transformations. In *Proc. of 27th International Conference on Very Large Data Bases (VLDB'01)*, Rome, Italy, September 2001.
- [FTS00] Mary F. Fernandez, Wang Chiew Tan, and Dan Suciu. SilkRoute: trading between relations and XML. *Computer Networks*, 33(1-6):723–745, 2000.
- [Gra04] Jim Gray. Distributed computing economics. In A. Herbert and K. Sparck Jones, editors, *Computer Systems Theory, Technology, and Applications, A Tribute to Roger Needham*, pages 93–101. Springer, 2004.
- [HHY⁺01] R.J. Millerand M.A. Hernández, L.M. Haas, L. Yan, C.T.H. Ho, R. Fagin, and L. Popa. The clio project: Managing heterogeneity. *SIGMOD Record*, 30(1), March 2001.
- [HT] Tony Hey and Anne Trefethen. The data deluge: An e-science perspective. http://www.rcuk.ac.uk/escience/documents/report_datadeluge.pdf referenced 20 July 2005.
- [Les] Michael Lesk. archiv.twoday.net/stories/337419/ referenced 22 July 2005.
- [LMM01] B. Ludäscher, R. Marciano, and R. Moore. Towards self-validating knowledge-based archives. In *11th Workshop on Research Issues in Data Engineering (RIDE)*, Heidelberg, Germany. IEEE Computer Society, April 2001.

- [OAI] Reference model for an open archival information system (oais). CCSDS 650.-B-1. Blue Book. Issue 1. Washington D.C. January 2002
<http://www.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf>.
- [PB03] R.A. Pottinger. and P.A. Bernstein. Merging models based on given correspondences. In *VLDB*, 2003.
- [SED] Lincoln D. Stein, Sean Eddy, and Robin Dowell. Distributed Sequence Annotation System (DAS). <http://www.biodas.org/documents/spec.html>.
- [SM03] Martin Szomszor and Luc Moreau. Recording and reasoning over data provenance in web and grid services. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 63–620. Springer, LNCS 2888, 2003.
- [Wad] P. Wadler. The links project.
<http://homepages.inf.ed.ac.uk/wadler/papers/links/links-blurb.pdf>.
- [XML] XML Schema Part 1: Structures Second Edition. <http://www.w3.org/TR/xmlschema-1/>.
- [XQu] XQuery 1.0: An XML Query Language. <http://www.w3.org/TR/xquery/>.