Report on the First International Workshop on Database Preservation (PresDB'07)

Vassilis Christophides
Institute of Computer Science
Foundation for Research and Technology-Hellas
P.O. Box 1385, GR-711 10 Heraklion, Greece
+30 2810 391628
christop@ics.forth.gr

Peter Buneman
School of Informatics and Digital Curation Centre,
University of Edinburgh, UK
Crichton Street, Edinburgh EH8 9LE
+44 131 650 5133
opb@inf.ed.ac.uk

1. INTRODUCTION

The need to preserve scientific, scholarly and cultural data has long been recognized. These data sets are valuable and many of them are either impossible to reproduce (e.g. climate and demographic data) or can only be recovered at enormous costs (e.g. data from high energy physics experiments). While substantial investment has been made in archiving and preserving conventional forms of these objects, such as documents, images and numerical data in some file format, the need to preserve entire *databases* has only recently emerged. Databases differ from fixed digital objects studied in the past, in that they change over time, they have internal structure, and they include schemas and integrity constraints, which are basic for the *current* and *future interpretation* of the data. Increasingly, database technology is being used in the storage of large numerical scientific data sets.

There is another use of databases for which preservation is equally important. Nowadays, nearly every on-line reference work, dictionary and gazetteer benefits from some form of database management support. These *curated* databases represent a huge investment of human effort, and they are becoming increasingly important in various scientific disciplines. The field of molecular biology alone boasts hundreds of curated databases. Archivists of both conventional (paper) and scientific data are also developing curated databases as catalogs for their holdings.

Despite the fact that databases are now central to scientific research and scholarship, very little thought has been devoted to their preservation. Databases are usually managed in central data centers and rely on the continued functioning of complex data management software as well as the funding for those centers. Existing preservation techniques for fixed digital objects are not suited for databases, thus some of our most critical digital assets are endangered – both economically and technically – in the long term.

The First International Workshop on Database Preservation[1] (PresDB'07) was organized by the UK Digital Curation Centre[2] (DCC) and held at the National e-Science Centre's e-Science Institute in Edinburgh, Scotland, on March 23, 2007. The goal of the workshop was to identify new technical, economic and legal issues arising in database preservation. The target audiences were researchers and practitioners working in the areas of databases, libraries and archives. Although announced at short notice, the event attracted 40 participants from 9 different countries in Europe and USA, and we believe that it was an important step in coordinating future research activities in this challenging topic. The eight invited talks and thirteen short presentations at the workshop[3] were divided into four sessions, which we summarize.

2. Session 1: Computer Scientists' Perspective on Database Preservation

This session focused on logical and physical aspects of information preservation related to the reliability and authenticity of long-term digital storage systems. All speakers recognized the scientific and technical challenges when preserving even conventional forms of digital objects such as multimedia documents.

Giorgos Flouris (*Istituto della Scienza e delle Tecnologie della Informazione, CNR, Pisa, Italy*) described a logic-based perspective on information preservation, which would allow a formal definition of the preservation problem as well as a characterization of desirable properties of existing and future preservation methods. To handle both the static and the dynamic aspects of digital preservation, he proposed not the preservation of the digital object itself (e.g. a database) but a set of related properties, such as answers to queries, which could include quality-related information (e.g. provenance) regarding the content of the object itself. Moreover, he argued that data should

[1] PresDB'07 is an informal workshop organized by a small executive committee (P. Buneman, V. Christophides, H. Müller, B. Ludaescher, C. Rusbridge, W.C. Tan, K. Thibodeau).

[2] www.dcc.ac.uk

[3] All presentations are available from the workshop website at homepages.inf.ed.ac.uk/hmueller/presdb07

be separated from the underlying community knowledge, which provides meaning. He finally presented useful connections of information preservation with the related fields of belief change/revision and ontology evolution.

Mema Roussopoulos (*Harvard University, USA and FORTH-ICS, Greece)* examined several threats to long-lived data from an end-to-end perspective. Not only hardware and software faults but also faults due to humans and organizations require consideration. She presented a simple model of long-term storage failures that helps us to reason about various strategies for addressing some of these threats. In particular, the most important strategies for increasing the reliability of long-term storage are: quickly detecting latent faults, automating fault repair to make it cheaper and faster, and increasing the independence of data replicas.

David Rosenthal (*Stanford University Libraries, USA*), drew on the experience of the LOCKSS system to highlight the economical risks when designing large-scale data preservation systems. As the size of the data to be preserved grows, the per-replica cost comes to dominate all other costs in the system. Minimizing the number of replicas needed to assure adequate preservation becomes the dominant design goal. At the petabyte scale of today's e-science databases, a single replica may cost over a million dollars. The presentation pointed out two basic questions: (a) How well prepared are we to take rational investment decisions about systems in this area? (b) Can we specify an acceptable risk of loss and from that determine how many replicas, how independent and how frequently audited we need? Some of the gaps identified by this talk included: the need for better specification and characterization of media performance (recent papers show that much of what we believe about disk reliability is wrong), better models of the possible threats (the most frequent cause of data loss at large sites is operator error), better models of fault tolerance (recent papers show that both RAID and Byzantine Fault Tolerance are inappropriate models), and better ways of formulating the relationship between a preservation service and its customers.

David Gross-Amblard (*Le2i Lab., Université de Bourgogne, France*) elaborated on the new threats (theft or information leakage) on Digital Archives given that purely digital objects are vulnerable to forgery and copying. In particular, when several production and preservation locations are involved, an inexperienced data producer may try to submit fake data (e.g. not produced by the correct device); an archivist may accidentally alter data e.g. by format mismatch; and malicious users could insert fake records or even fake an entire archive from available archives. To this end, he presented several digital watermarking techniques that allow us to preserve authenticity and provenance of database records. Finally, he highlighted the strengths of watermarking techniques for preserving the accuracy of numeric database queries to an acceptable degree.

3. Session 2: Brainstorming Session

The main questions addressed by the speakers were:

- How do we keep, at acceptable cost, archived databases readable and usable in the long?
- How do we separate the data from a specific database management environment?
- How can we preserve the original data semantics and structure?
- How can we preserve authenticity and provenance of databases?
- How can we preserve data while it continues to evolve?
- How can we have efficient preservation frameworks, while retaining the ability to query different database versions?
- How can multi-user online access be provided to archived databases containing terabytes of data?
- Can we move from a centralized to a distributed, redundant model of database preservation?

Peter Buneman's (*School of Informatics and Digital Curation Centre, University of Edinburgh, UK*) talk "Why Current Database Technology Does not Support Preservation" noted that existing DBMS history support (e.g. ORACLE 10g and Microsoft) is mainly designed for correction/recovery and "flashback" queries, which raise serious efficiency concerns when very long transaction sequences are rolled back. Even if we ignore the issues of long-term preservation of complex database software, current systems provide limited support for temporal/longitudinal queries and temporal references (citations). More importantly, although curated databases use database technology, the contents of the database seldom includes all the data of interest. This is frequently held in structures external to the database or is embedded in the applications programs. In addition, the database schema may only provide weak support for the interpretation of the data. One could hope that better database design would alleviate some of these problems, but that is likely to be wishful thinking. Preservation of scientific databases requires much more than preservation of the tables and schema; it requires a unified approach to publishing and archiving.

George Papastefanatos (*National Technical University of Athens, Greece*) stressed the relationship of a database with its surrounding applications. Clearly, the timespan of a preserved database is much larger than that of an application deployed over it, but applications do matter: (i) 50% of our effort goes into their maintenance, (ii) much of the semantics (constraint checking and business rules) is hard-coded in the applications for ease of programming, and (iii) despite any changes in the underlying infrastructure (database schema, development platform, etc.), the organization still needs certain applications to continue indefinitely. In fact, in public sector data, some applications remain in use for decades, despite technological advances. Thus, the main challenge is to study common formalisms to express the structure and semantics of a database-centric environment as well as to devise frameworks for responding to change. In particular, he sketched a self-monitoring, auto-regulating, and self-repairing approach for evolving database-centric systems, in which queries and views are adapted on demand to events that alter the underlying data structure and semantics.

Stefan Brandl (*CSP GmbH & Co. KG, München, Germany*) presented the Chronos experience of building highly scalable long-term archives of relational databases for companies, organizations, science data centers, and public archives. The dependency of the data on original production environments is removed, but the original semantics, structure and integrity of the data is retained and preserved in an open Text/XML-based format that does not require any database management system to maintain, access and retrieve archive data. Chronos can collectively maintain any number of database archives, and provides easy web-based access to any number of users from any location. Furthermore, Chronos can be seamlessly integrated with the life cycle management of production systems: although the database schema of a production system may evolve over time, it can continuously and incrementally extract data subsets and yet ensure coherent and collective accessibility and manageability for all archived data. This is possible because of Chronos' ability to detect, describe and manage the semantic and structural changes in the production database schema between any two subsequent executions of the archiving process.

The Norman Swindells (*CEng, FIMMM Ferroday Ltd, Birkenhead, UK*) pointed out that maintaining data for longer than the lifetime of the DBMS software is a problem that the manufacturing industry has already faced. He reported lessons learned from STEP/EXPRESS and PLIB standards in representing complex manufacturing data, e.g. aircrafts and automobiles. Finally, Ulf Andersson (*AstraZeneca, Sweden*) presented case studies from long term preservation applications in the pharmaceutical industry.

4. Session 3: An Archivist's/Librarian's perspective on Database Preservation

This session consisted of four presentations addressing changes in modern scientific and scholarly work practices, as well as, related technical challenges involved in the life-cycle of preservation of scientific databases (ingestion, management, storage and access).

John A. Kunze (*University of California, Office of the President, USA*) examined the changing nature of digital object citations especially in the light of the recent explosion of e-science. While traditional citation guidelines have been sufficient to create stable references for printed materials held in the world's libraries and archives, common citation practices have not yet appeared for digital objects that change frequently, come in a variety of formats, and themselves consist of a hierarchy of citable objects. The absence of such practices is keenly felt in scientific research that relies on long-term access to large databases. The speaker stressed the need for an underlying usage model which should be acceptable by data users, producers, publishers, and archivists. The model should ideally encompass a variety of citation needs and synthesize prior discipline-specific efforts at data description in biomedicine, political and social science, astronomy, geography, etc. Citations should include persistent identifiers usable with widely available software to gain access to cited works.

Kevin Ashley (*University of London Computer Centre, UK*) expressed practical concerns regarding "dirty" or absent data, especially when databases to be preserved are far removed from their creation (unlike curated databases). Most existing preservation tools, data models and representation methods assume that data is clean, and conforms to some design (integers are really integers, rather than random bits of text, for instance). For many applications this is true, but when we are trying to preserve database records, the flaws they contain are important. When bad data is used to inform important decisions, we should record this process. We may need to know not just what malformed data existed, but how it effected the environment in which it was used. The speaker advocated a framework for dealing with imperfect data by preserving the informative nature of errors yet allowing their effective manipulation. Ideally, this would also represent missing or absent data, a problem which appears in the social and experimental sciences but which tends to be treated in many different domain-specific ways. Finally, he noted that timed embargoes are often required on preserved data: one may want to preserve it immediately but delay access for a specified period. New control policies are needed for this.

Bill Roberts (*Tessella - Scientific Software Solutions, UK*) reported success stories in the area of scientific data preservation, and identified "easy" and "hard" parts of the problem. In particular, much of the success in the preservation of experimental nuclear fusion data (since the mid 1980s) can be assigned to: (a) having clear responsibilities for maintaining the data, (b) the continuing existence of the organization that created the data and controls the data formats, and (c) planning for widespread access to the from the start. By contrast in, the oil and pharmaceutical companies there remain a number of difficulties, arising from the diversity of data and processes (a) important contributions are made by a large number of groups, each using database systems designed to meet their individual needs (b) data formats are often determined by the suppliers of specialized scientific instruments (c) interpretation of the data may depend on complex algorithms (e.g. database stored procedures, visualization software etc.) and these become obsolete rapidly and are hard to migrate (d) scientific techniques, processes and instruments used to produce the data also evolve rapidly (e) changes in organizational structure (e.g. company take-overs, mergers) lead to a complex data history.

Michael Lesk (*Department of Library and Information Science, Rutgers University, USA)* encouraged scholars to conduct a new kind of research. Exploiting old information becomes secondary, since in times past there was not much to exploit and not much to gain from reanalyzing data. Today meta-analysis of old studies and data mining of repositories of automatically collected information is a productive source of new science and scholarship, but it is still not taught or properly rewarded. The speaker highlighted the need for a new set of values in the scholarly community if we are to get adequate resources devoted to data curation.

5. Session 4: Brainstorming Session

The main questions addressed by the speakers were:

- What are the salient features of a database that should be preserved?
- What are the different stages in the database preservation's life cycle?
- What documentation is preserved together with a database, and in what format?
- What are the legal encumbrances on database preservation?
- What can be learned from traditional archival appraisal for the selection of databases for preservation?
- To what extent can the preservation strategies, and procedural policies developed by archivists be adapted for databases?

W. Christopher Lenhardt (*Center for International Earth Science Information Network, Columbia University* and *NASA Socioeconomic Data and Applications Center*) highlighted technical and organizational challenges in building trusted digital repositories. Katerina Tsakona (*Department of Computer Science, University of Crete, Greece*) detailed legal encumbrances on database preservation and in particular intellectual property rights management. Whether actions such as copying, reproduction or re-creating the entire database and/or its constituents, are simply lawful or not, will always depend upon the permission of the appropriate parties, the right-holders, whether they are known or not. Luís Faria (*National Archives of Portugal*) presented the functional specifications and conceptual models of the RODA project. RODA relies on an abstraction layer (XML-based) for capturing the DB structure and information, which is independent from the underlying platform and RDBMS. Dirk Roorda (*Data Archiving and Networking Services, Netherlands*) presented the MIXED project addressing digital preservation of spreadsheets and databases in social sciences, the arts and the humanities. One of the practical assumptions made in this project is that only the semantic core of the data is archived. Loosely speaking, it is the meaning of the data, independent of the associated presentation and actions. For example: a table of outcomes of censuses in the Netherlands during the 19^{th} century is mainly interesting for its values, and not for the font family (presentation), or on the process employed to record those values. However, in a database context, ignoring business rules, expressed in constraints and triggers, can lead to loss of useful information. Finally, Seamus Ross (*University of Glasgow, UK*) presented interesting uses cases when preserving performing arts databases; Jonathan Bard (*School of Biomedical Sciences, University of Edinburgh, UK*) reported the experience of the European Radiobiological Archives; a collection of hugely valuable and irreproducible radiation studies; and Rolf Lang (*Landesarchiv Baden-Württemberg, Germany*) highlighted practical concerns regarding existing technology support for preserving databases.