DATABASES IN HISTORICAL RESEARCH

Databases in Historical Research

Theory, Methods and Applications

Charles Harvey and Jon Press

palgrave
macmillan

© Charles Harvey and Jon Press 1996

Softcover reprint of the hardcover 1st edition 1996 978-0-333-56843-9

All rights reserved. No reproduction, copy or transmission of this publication may be made without written permission.

No paragraph of this publication may be reproduced, copied or transmitted save with written permission or in accordance with the provisions of the Copyright, Designs and Patents Act 1988, or under the terms of any licence permitting limited copying issued by the Copyright Licensing Agency, 90 Tottenham Court Road, London W1T 4LP.

Any person who does any unauthorised act in relation to this publication may be liable to criminal prosecution and civil claims for damages.

The authors have asserted their rights to be identified as the authors of this work in accordance with the Copyright, Designs and Patents Act 1988.

Published by

PALGRAVE MACMILLAN

Houndmills, Basingstoke, Hampshire RG21 6XS and

175 Fifth Avenue, New York, N. Y. 10010

Companies and representatives throughout the world

PALGRAVE MACMILLAN is the global academic imprint of the Palgrave Macmillan division of St. Martin's Press, LLC and of Palgrave Macmillan Ltd. Macmillan® is a registered trademark in the United States, United Kingdom and other countries. Palgrave is a registered trademark in the European Union and other countries.

ISBN 978-0-333-56844-6

ISBN 978-1-349-24392-1 (eBook)

DOI 10.1007/978-1-349-24392-1

This book is printed on paper suitable for recycling and made from fully managed and sustained forest sources.

A catalogue record for this book is available from the British Library.

Contents

<i>List of Figures</i>	viii
<i>Preface</i>	xi
1. <i>Databases in Historical Research</i>	1
1.1 The Processes of Historical Research	2
1.2 Databases in Historical Research	5
1.3 Types of Historical Databases	10
Case Study A: Investigating Regional Economies: The Gloucester Portbooks Database (Peter Wakelin and David Hussey)	14
2. <i>Database Concepts and Terminology</i>	22
2.1 Database Concepts	22
2.2 Evolution of Database Systems	25
2.3 Features of Database Systems	33
2.4 Conclusion	39
Case Study B: The Register of Music in London Newspapers, 1660–1800 (Catherine Harbor)	40
3. <i>Database Management Software</i>	47
3.1 Flat File Managers	49
3.2 Database Management Systems	50
3.3 Textual Information Management Systems	53
3.4 Integrated Systems	61
3.5 Multimedia and Hypermedia	62
3.6 Conclusion	67
Case Study C: Hypermedia Database Management Systems: Microcosm as a Research Tool (Frank Colson)	69
4. <i>The Database Project Life-Cycle</i>	73
4.1 Project Management	73
4.2 Project Planning	76
4.3 Source Analysis	80
4.4 Database Design	82

4.5	Database Implementation	84
4.6	Data Entry and Validation	85
4.7	Database Administration	94
4.8	Conclusion	97
	Case Study D: Creating a Machine-Readable Version of the 1881 Census of England and Wales (Matthew Woollard)	98
5.	<i>Database Design and Implementation</i>	102
5.1	Approaches to Database Design	102
5.2	Entity-Relationship Modelling	103
5.3	Relational Data Analysis	119
5.4	Creating a Database	130
5.5	Conclusion	137
	Case Study E: Indicators of Regional Economic Disparity: The Geography of Economic Distress in Britain before 1914 (David Gilbert and Humphrey Southall)	140
6.	<i>Information Retrieval</i>	147
6.1	Query-by-Example	147
6.2	SQL	153
6.3	Querying a Textbase	166
6.4	Reports	176
6.5	Conclusion	184
	Case Study F: Analysing Social Structure and Political Behaviour using Poll and Rate Books: The Westminster Historical Database (Edmund Green)	185
7.	<i>Source-Oriented Database Systems</i>	188
7.1	Source-Oriented Data Processing	190
7.2	Multimedia and Hypermedia Databases	198
7.3	Object-Oriented Databases	203
7.4	Artificial Intelligence, Expert Systems and Knowledge Bases	208
7.5	Conclusion	217
	Case Study G: Community Reconstruction and the Viana Do Castelo Database (Jean Colson)	218
8.	<i>Coding and Record Linkage</i>	222
8.1	Coding	222

8.2 Record Linkage	234
8.3 Conclusion	252
Case Study H: Researching the Population History of England (Kevin Schürer)	253
9. Conclusion: Databases and the Future of Historical Computing	257
<i>Notes and References</i>	261
<i>Glossary</i>	283
<i>Bibliography</i>	297
<i>Index</i>	320

List of Figures

1.1	The processes of historical research	3
1.2	Computing and the processes of historical research	7
1.3	Centrality of databases to historical computing	9
A.1	A typical page from the Port Books	15
A.2	Sample data entry form	17
A.3	Data item list for the Portbooks Database	18
A.4	Recorded downstream salt shipments from Gloucester, 1660–1760	21
2.1	Database terminology	25
2.2	Data processing: data and meaningful information	26
2.3	Applications-centred approach to data processing	27
2.4	The database-centred approach	28
2.5	Representation of the hierarchical model	30
2.6	Representation of the network model	31
2.7	A query in Structured Query Language (SQL)	36
B.1	Numbers of references, 1660–1720	46
3.1	The ORACLE relational database system	53
3.2	Other types of text retrieval systems	59
4.1	Stages in project development	76
4.2	Music-related items in London newspapers, 1660–1745	79
4.3	Newspaper content analysis	82
4.4	Schemas	83
4.5	Stages in data preparation and entry	88
5.1	One-to-one relationship	104
5.2	One-to-many relationship	104
5.3	Many-to-many relationship	104
5.4	Six-stage approach to entity-relationship modelling	109
5.5	Initial entity-relationship diagram	111
5.6	Redundant relationships	113
5.7	Redundant link entities	113
5.8	Tables before dissolution of many-to-many relationship	114
5.9	Tables after dissolution of many-to-many relationship	115
5.10	Final entity-relationship diagram	117
5.11	Entity description form for the entity Membership	118
5.12	Normalisation	120
5.13	Removal of repeating groups	121
5.14	Removal of partial dependency	122

5.15	Removal of indirect (transitive) dependency	123
5.16	Unnormalised data item list	125
5.17	Transformation to first normal form through removal of repeating groups	126
5.18	Dependency diagram for the Seaman table	127
5.19	Transformation to second normal form through removal of partial dependency	127
5.20	Dependency diagram for the Seaman_Voyage table	129
5.21	Transformation to third normal form through removal of indirect (transitive) dependency	129
5.22	Microsoft Access: the database window	135
5.23	Microsoft Access: table definition	136
5.24	Microsoft Access: form definition	138
E.1	Contents of QMW Labour Markets Database	142
E.2	A screen from the Labour Markets Database	146
6.1	Guidelines for interrogating a database or textbase	148
6.2	Microsoft Access: The Indgroup database window	150
6.3	Microsoft Access: The Seat form	150
6.4	Microsoft Access: single table query – design view	151
6.5	Microsoft Access: single table query – datasheet view	151
6.6	Microsoft Access: using OR in querying	152
6.7	Microsoft Access: query form for the multitable query	153
6.8	Microsoft Access: result of the multitable query	153
6.9	Using SQL in Microsoft Access	166
6.10	SQL*TextRetrieval: using the thesaurus	169
6.11	SQL*TextRetrieval: the synonym ring for 'Handel'	170
6.12	SQL*TextRetrieval: other thesaurus links relating to 'Handel'	170
6.13	SQL*TextRetrieval: using LIST TABLES	172
6.14	SQL*TextRetrieval: the hitlist resulting from a query	173
6.15	SQL*TextRetrieval: narrowing down the search	174
6.16	Steps in developing a report	177
6.17	Sample printout of the completed report	181
6.18	Microsoft Access: prompt screens in the Report Wizard	182
6.19	Microsoft Access: the design view of the report	183
6.20	Microsoft Access: an extract from the printed report	183
7.1	Relationship between the user and κλειώ	193
7.2	κλειώ – user view	194
7.3	κλειώ – technical description	195
7.4	Lists of datasets which may be mapped on to regions	201
7.5	Mapping unemployment data	202
7.6	Generating graphs from data held in the underlying database table	203

7.7	Object orientation concepts	206
7.8	Components of an expert system	209
7.9	Confidence factors	211
7.10	Strength of set membership	211
8.1	The United Kingdom Standard Industrial Classification System (extract)	226
8.2	Soundex coding	230
8.3	Examples of Soundex coding	230
8.4	Weights for linkage keys	245
8.5	Point scoring of algorithms 1–30	246
8.6	Algorithms ranked by relative confidence	247
8.7	SQL record linkage query	248
8.8	Linkage rates achieved	249
8.9	Regression residuals	250

Preface

This book is intended as a contribution to historical methodology: to the body of literature explicitly concerned with the ways and means of creating and representing historical knowledge. Its starting point is the observation that the construction of historical databases has become in recent years a widespread activity consuming resources on a large scale. A recent guide published by the Association for History and Computing, though 'far from being an exhaustive or definitive inventory', gives details of 376 projects carried out in the United Kingdom that have involved the creation of machine-readable data files.¹ These projects have been carried out on a wide variety of subjects covering all historical periods and many different types of source material. It is true that there has been a concentration on the nineteenth century, and that census, poll book and parish register data have predominated, but the fact remains that database-centred research is generally perceived to be applicable to all fields of historical enquiry.

The particular appeal of database-centred research is not difficult to understand. Data gathering, organisation, sorting and searching, are tasks routinely carried out by historians. Each is made less burdensome through the use of powerful computers and database management software, and the time and effort saved in this way have served to inspire ambition: historians at the leading edge nowadays conceive of very large databases, often incorporating diverse forms of data from many sources. The recurrent idea is that rich collections of data of this kind should form a collective resource, serving the needs of many scholars and many projects over many generations. This conception of databases as an enduring historical resource is expressed repeatedly in the landmark *History and Computing* set of three volumes published by Manchester University Press between 1987 and 1990. These contain versions of 113 of the papers delivered at the first three major conferences of the Association for History and Computing held in the United Kingdom. Databases and database technology are the main subjects of 53 papers and a minor subject of a further 16.²

The use of databases in historical research is set to increase to the point where a basic knowledge of database systems will be regarded as an essential skill for all professionally trained historians. The value of database systems is already widely understood, and as knowledge and experience increase so too will the number of database projects. An important factor encouraging the launch of fresh projects is the es-

entially non-threatening nature of database technology. Historians must create and interrogate sets of data whether or not these take the form of a computer database. Thus, in using a database system, the historian is not adopting a fresh research paradigm, but merely a technology which, on the face of it at least, supports traditional methods of research and analysis.

The rising popularity of database-centred research, none the less, poses very real challenges for the historical community. New skills, methods, standards and thinking are required if maximum advantage is to be gained from the technology. Historians must have a sufficiently deep technical understanding of database systems to select the right hardware and software for a project, to design databases that effectively satisfy the requirements of a project, and to construct, manage and exploit to the full the potentialities of project databases. They must also understand how databases fit within the bigger picture of computer-assisted historical research. Database design, construction and interrogation form just one of a number of sets of related processes that constitute *historical computing*. It is essential that historians working with databases should keep in mind the project as a whole and the interaction between processes. This indicates a need to think seriously about project management, and about standards of quality and transportability. There is little point in creating a database that, through a failure to conform to agreed standards, cannot satisfy the requirements of potential users.

In this book, we provide an introduction to database systems that will be of interest and immediate value to anyone undertaking a database-centred historical research project. We believe that a sound knowledge of database systems can assist the work of all sorts of historian: professional academics, postgraduate and undergraduate students, and amateurs working on regional and local projects. Our approach is to explore the subject within the broad framework of historical computing. We consider in depth the topics and issues which have most bearing upon the processes of historical research. Some of these are purely technical. Others are methodological. Our intention is to offer an account of database systems in historical research that will enable the reader to avoid costly mistakes and realise the full potential of available resources.

In Chapter 1, we consider the nature and processes of historical research, and the ways in which computers may be used to assist the historian obtain better results. The centrality of database systems to historical computing is explained. The main technical differences between database projects and types of database are established. On completing this chapter, the reader should have in mind the reasons for

the importance of databases and database systems to computer-assisted historical research.

The next two chapters are fundamental. Chapter 2 is an introductory survey of database concepts, terminology and methods. This provides the conceptual underpinning which is needed before we can look in any depth at historical research projects. Though the discussion is necessarily technical, important terms and concepts are simply explained and illustrated diagrammatically. In Chapter 3, attention is given to the differing types of software available, from relatively simple file managers to powerful and complex relational database management systems, textual information managers, and multimedia systems. We note the dominance of relational systems in the marketplace, but stress that in recent years the development of database technology has progressed in new directions as users have demanded greater choice and flexibility. This is a theme to which we return in Chapter 7, where we consider some of the strategies and approaches which have been pursued by the critics of relational systems.

Chapters 4–6 are concerned with the design, construction and querying of databases. The planning, specification and management of database research projects is examined in Chapter 4. It is argued that effective planning and review procedures are critical to the success of large projects. A number of mundane but important issues relating to the development of a database are also discussed, including data preparation, entry and validation. In Chapter 5, we provide a step-by-step guide to two important approaches to database design – entity-relationship modelling and relational data analysis – and work through the actual creation of a database. Historians must be skilled in retrieving information from databases, and in Chapter 6 the main methods of forming queries and accessing databases are examined. Database reports and the presentation of information are also considered.

The example databases used in these chapters are implemented using two leading database management systems, ORACLE and Microsoft Access. ORACLE provides a very powerful set of programs which is available on a wide range of platforms from mainframes to personal computers. It is widely available on university networks. It uses SQL (Structured Query Language), the industry standard for creating, managing and querying databases, and has added extensions for managing large quantities of unstructured text, and displaying graphically the results of queries. Access is a much later arrival on the scene, but has rapidly established a dominant position amongst PC-based database management systems. Access is an up-to-date Windows product which is both powerful and easy to use. It provides a user-friendly form of information retrieval called query-by-example, and also offers SQL

facilities for the benefit of more experienced users.

Chapter 7 reviews the concept of source-oriented data processing and the ambitious κλειω (kleio) project, and the advance of multimedia and hypermedia, object orientation, knowledge bases and expert systems. In Chapter 8, attention is given to two key issues in historical computing. The first concerns the need for historians to introduce codes into databases for reasons of order, classification and analysis. The second concerns the use of database systems as tools for record linkage: for drawing together disparate pieces of information relating to a single entity, most often a person.

The final chapter sets out the key points which must be borne in mind if the tremendous potential of database-centred research is to be realised. A select bibliography and a glossary of concepts and technical terms is also provided.

A distinctive feature of this book is the presentation of case studies at the close of each chapter. These are the work of project leaders and others closely acquainted with major database-centred research or software development projects: they are Peter Wakelin (CADW) and David Hussey (University of Wolverhampton); Catherine Harbor (Royal Holloway, University of London); Frank Colson (University of Southampton); Matthew Woollard (University of the West of England, Bristol); David Gilbert (Royal Holloway, University of London) and Humphrey Southall (Queen Mary & Westfield College, University of London); Edmund Green (Royal Holloway, University of London); Jean Colson (University of Southampton); and Kevin Schürer (ESRC Data Archive, University of Essex). The cases bring the book to life in a way that would otherwise be impossible. They are referenced throughout the main body of the text to illustrate key points. As stand-alone essays, they are records of achievement which we hope will inspire readers, suggest research possibilities, and highlight the main advantages of database systems and the methods of historical computing. We are most grateful to their authors for the time and trouble taken in preparing them.

We are also indebted to many other friends, colleagues and associates for their assistance in preparing this book for publication. They have given generously of their time, explaining their database projects in detail and helping us to clarify our thoughts on various matters. In particular, we would wish to record our thanks to Peter Denley (Queen Mary & Westfield College, University of London) and Deian Hopkin (London Guildhall University), the founders of the Association for History and Computing; Steve Baskerville (University of Hull); Leen Breure (Rijksuniversiteit Utrecht); Virginia Davis (Queen Mary & Westfield College, University of London); David Devlin (Bath College of Higher

Education); David Doulton (University of Southampton); Mike Goodland (University of Kingston); Philip Hartland (Roehampton Institute); Dave Hassall (Bath College of Higher Education); Noel Heather (Royal Holloway, University of London); Eddie Higgs (Welcome Unit for the History of Medicine, Oxford); Jack Hoare (Genealogical Society of Utah); Rosamund McGuinness (Royal Holloway, University of London); Roger Middleton (University of Bristol); Bob Morris (University of Edinburgh); Burkhard Pöttler (University of Graz); Manfred Thaller (University of Göttingen); Richard Trainor (University of Glasgow); John Turner (Royal Holloway, University of London); and Peter Wardley (University of the West of England, Bristol). Gwyn Price and Alec Gray (University of Wales College of Cardiff) kindly let us see a pre-publication copy of their paper on object-oriented databases and their application to historical research. Catherine Harbor and Matthew Woollard undertook a critical reading of the entire typescript. As a result of their collective efforts, we have been able to make substantial improvements in technical accuracy and clarity of explanation. As always, however, any errors or deficiencies which remain are ours alone.

We should also like to record our gratitude to the Scouloudi Foundation (formerly the Twenty-Seven Foundation) and the Research Committee of Bath College of Higher Education, which kindly supported our research through the provision of funds for travel and other expenses. Parts of Chapters 5 and 7 previously appeared in the pages of *History and Computing*, and we are grateful to the editor and publishers for permission to reprint them here. Finally, we would like to thank the staff of Macmillan Press, and especially Vanessa Graham, Simon Winder, Margaret Bartley, Pauline Underwood and Keith Povey, for their consideration, patience and warm support throughout.

CHARLES HARVEY
JON PRESS

■ Note on Terminology and Text Styling

Throughout the book, the term 'table' is used rather than 'relation' to describe the two-dimensional files that comprise a relational database. Although purists might object that, strictly speaking, the latter term is more correct, its use can lead to confusion between 'relation' and 'relationship'. 'Record' is generally preferred to 'row', and 'field' to 'column', though both terms are used. To distinguish them from

the text, computer commands and printouts are in a Courier (monospaced) font. Additionally, for ease of identification, the names of tables and entities are in a bold type, and fieldnames in italics (though this would not be required in practice). All trade marks and registered names are acknowledged.