Principles of Data Management

Facilitating Information Sharing

Keith Gordon
Contents

Figures and tables xi
About the author xv
Foreword xvii
Glossary xix
Preface xxv
Introduction xxix

1 Data and the Enterprise 1
   Information is a key business resource 1
   The relationship between information and data 2
   The importance of the quality of data 4
   The common problems with data 5
   An enterprise-wide view of data 7
   Managing data is a business issue 8
   Summary 10

2 Database Development 11
   The database architecture of an information system 11
   An overview of the database development process 16
   Conceptual data modelling (from a project-level perspective) 20
   Relational data analysis 33
   The roles of a data model 49
   Physical database design 50
   Summary 53

3 What is Data Management? 54
   The problems encountered without data management 54
   Data management responsibilities 57
   Roles within data management 60
   The benefits of data management 62
   The relationship between data management and enterprise architecture 62
   Summary 64
## Contents

<table>
<thead>
<tr>
<th>Chapter</th>
<th>Title</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Corporate Data Modelling</td>
<td>65-68</td>
</tr>
<tr>
<td></td>
<td>Why develop a corporate data model?</td>
<td>65</td>
</tr>
<tr>
<td></td>
<td>More data modelling concepts</td>
<td>66</td>
</tr>
<tr>
<td></td>
<td>The nature of a corporate data model</td>
<td>72</td>
</tr>
<tr>
<td></td>
<td>How to develop a corporate data model</td>
<td>74</td>
</tr>
<tr>
<td></td>
<td>Corporate data model principles</td>
<td>78</td>
</tr>
<tr>
<td></td>
<td>A final thought</td>
<td>82</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>83</td>
</tr>
<tr>
<td>5</td>
<td>Data Definition and Naming Conventions</td>
<td>84-88</td>
</tr>
<tr>
<td></td>
<td>The elements of a data definition</td>
<td>84</td>
</tr>
<tr>
<td></td>
<td>Data naming conventions</td>
<td>88</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>90</td>
</tr>
<tr>
<td>6</td>
<td>Metadata</td>
<td>91-93</td>
</tr>
<tr>
<td></td>
<td>What is metadata?</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Metadata for data management</td>
<td>91</td>
</tr>
<tr>
<td></td>
<td>Metadata for content management</td>
<td>92</td>
</tr>
<tr>
<td></td>
<td>Metadata for describing data values</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>93</td>
</tr>
<tr>
<td>7</td>
<td>Data Quality</td>
<td>95-101</td>
</tr>
<tr>
<td></td>
<td>What is data quality?</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>Issues associated with poor-quality data</td>
<td>95</td>
</tr>
<tr>
<td></td>
<td>The causes of poor-quality data</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>The dimensions of data quality</td>
<td>97</td>
</tr>
<tr>
<td></td>
<td>Data model quality</td>
<td>98</td>
</tr>
<tr>
<td></td>
<td>Improving data quality</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>101</td>
</tr>
<tr>
<td>8</td>
<td>Data Accessibility</td>
<td>102-111</td>
</tr>
<tr>
<td></td>
<td>Data security</td>
<td>102</td>
</tr>
<tr>
<td></td>
<td>Data integrity</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>Data recovery</td>
<td>109</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>111</td>
</tr>
<tr>
<td>9</td>
<td>Database Administration</td>
<td>112-115</td>
</tr>
<tr>
<td></td>
<td>Database administration responsibilities</td>
<td>112</td>
</tr>
<tr>
<td></td>
<td>Performance monitoring and tuning</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Summary</td>
<td>115</td>
</tr>
<tr>
<td>10</td>
<td>Repository Administration</td>
<td>116-122</td>
</tr>
<tr>
<td></td>
<td>Repositories, data dictionaries, encyclopaedias, catalogs and directories</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>Repository features</td>
<td>118</td>
</tr>
<tr>
<td></td>
<td>The repository as a centralised source of information</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>Metadata models</td>
<td>122</td>
</tr>
</tbody>
</table>
Contents

Summary 122

11 The Management of Data Management 123
   Techniques and skills for data administration 123
   Techniques and skills for database administration 124
   Techniques and skills for repository administration 125
   The positioning of data management within the enterprise 125
   Summary 130

12 Industry Trends and their Effects on Data Management 131
   The use of packages 131
   Distributed data and databases 133
   Data warehousing and data mining 139
   Object orientation and databases 145
   Multimedia and databases 152
   Data and web technology 155
   Summary 158

A Comparison of Data Modelling Notations 159

B Hierarchical and Network Databases 169

C Generic Data Models 177

D An Example of a Data Naming Convention 183

E Metadata Models 195

F A Data Mining Example 201

G HTML and XML 205

H XML and Relational Databases 213
   References 219
   Further reading 221
   Index 223
Figures and tables

Figure 1.1 The relationship between data and information 4
Figure 2.1 A model of a database system 12
Figure 2.2 The three level schema architecture 15
Figure 2.3 A simplified view of the database development process 17
Figure 2.4 A conceptual data model diagram 18
Figure 2.5 A portion of an SQL create script 19
Figure 2.6 The EMPLOYEE entity 21
Figure 2.7 The attributes of the EMPLOYEE entity 22
Figure 2.8 The ADDRESS entity 23
Figure 2.9 The ‘resident at’ relationship 24
Figure 2.10 The QUALIFICATION and EMPLOYEE QUALIFICATION entities 26
Figure 2.11 The GRADE and EMPLOYEE GRADE entities 27
Figure 2.12 The DEPARTMENT and ASSIGNMENT entities 28
Figure 2.13 The one-to-one ‘managed by’ relationship 29
Figure 2.14 The many-to-many ‘managed by’ relationship 30
Figure 2.15 The resolution of the many-to-many ‘managed by’ relationship 30
Figure 2.16 The EMPLOYEE NEXT OF KIN entity 32
Figure 2.17 A relation shown as a table 34
Figure 2.18 The human resources paper record 36
Figure 2.19 The un-normalised EMPLOYEE ‘relation’ 37
Figure 2.20 The first normal form relations 39
Figure 2.21 Diagram of the first normal form relations 41
Figure 2.22 The second normal form relations 43
Figure 2.23 Diagram of the second normal form relations 44
Figure 2.24 The third normal form relations 46
Figure 2.25 Diagram of the third normal form relations 47
Figure 3.1 Data management activities 58
Figure 3.2 Data management deliverables 60
Figure 3.3 The business drives; data management steers 60
Figure 3.4 The relationship between data management and information management 61
Figure 4.1 The human resources conceptual data model developed in Chapter 2 67
Figure 4.2 Entity subtypes 68
Figure 4.3 An exclusive arc 70
Figure 4.4 Subtyping instead of using an exclusive arc 71
Figure 4.5 A first top-down starter model 76
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.6</td>
<td>A second top-down starter model</td>
<td>77</td>
</tr>
<tr>
<td>4.7</td>
<td>A third top-down starter model</td>
<td>77</td>
</tr>
<tr>
<td>4.8</td>
<td>The supply-chain model</td>
<td>79</td>
</tr>
<tr>
<td>4.9</td>
<td>The improved supply-chain model</td>
<td>80</td>
</tr>
<tr>
<td>4.10</td>
<td>More than one type of support?</td>
<td>80</td>
</tr>
<tr>
<td>4.11</td>
<td>The combined support model</td>
<td>81</td>
</tr>
<tr>
<td>4.13</td>
<td>The trade-off triangle</td>
<td>82</td>
</tr>
<tr>
<td>5.1</td>
<td>A data definition with validation criteria and valid operations</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>A data definition with valid values</td>
<td>87</td>
</tr>
<tr>
<td>5.3</td>
<td>An entity definition</td>
<td>88</td>
</tr>
<tr>
<td>7.1</td>
<td>The dimensions of data quality</td>
<td>98</td>
</tr>
<tr>
<td>7.2</td>
<td>The five dimensions of data model quality</td>
<td>99</td>
</tr>
<tr>
<td>7.3</td>
<td>Total Quality data Management methodology</td>
<td>101</td>
</tr>
<tr>
<td>8.1</td>
<td>Table privilege statements</td>
<td>103</td>
</tr>
<tr>
<td>8.2</td>
<td>A function privilege statement</td>
<td>103</td>
</tr>
<tr>
<td>8.3</td>
<td>A database object privilege statement</td>
<td>104</td>
</tr>
<tr>
<td>8.4</td>
<td>A view statement and an associated table privilege statement</td>
<td>104</td>
</tr>
<tr>
<td>8.5</td>
<td>A user-specific view statement and an associated table privilege statement</td>
<td>104</td>
</tr>
<tr>
<td>10.1</td>
<td>The role of directories or catalogs</td>
<td>116</td>
</tr>
<tr>
<td>10.2</td>
<td>The relationship between a CASE tool and its encyclopaedia or data dictionary</td>
<td>117</td>
</tr>
<tr>
<td>10.3</td>
<td>The architecture of a repository</td>
<td>118</td>
</tr>
<tr>
<td>10.4</td>
<td>The scope of a repository</td>
<td>119</td>
</tr>
<tr>
<td>10.5</td>
<td>The repository procurement process</td>
<td>120</td>
</tr>
<tr>
<td>10.6</td>
<td>A repository as a centralised source of information</td>
<td>121</td>
</tr>
<tr>
<td>11.1</td>
<td>The data administration skill set</td>
<td>123</td>
</tr>
<tr>
<td>11.2</td>
<td>Business-based data management</td>
<td>127</td>
</tr>
<tr>
<td>11.3</td>
<td>Independent data management in the IT/IS department</td>
<td>127</td>
</tr>
<tr>
<td>11.4</td>
<td>Data management within systems development</td>
<td>128</td>
</tr>
<tr>
<td>11.5</td>
<td>Dispersed IT/IS based data management</td>
<td>129</td>
</tr>
<tr>
<td>11.6</td>
<td>Distributed data management</td>
<td>129</td>
</tr>
<tr>
<td>12.1</td>
<td>Vertical fragmentation</td>
<td>136</td>
</tr>
<tr>
<td>12.2</td>
<td>Hybrid fragmentation</td>
<td>136</td>
</tr>
<tr>
<td>12.3</td>
<td>A typical data warehouse architecture</td>
<td>139</td>
</tr>
<tr>
<td>12.4</td>
<td>A multidimensional data model</td>
<td>141</td>
</tr>
<tr>
<td>12.5</td>
<td>A typical relational schema for a data warehouse</td>
<td>144</td>
</tr>
<tr>
<td>12.6</td>
<td>A simple conceptual data model</td>
<td>147</td>
</tr>
<tr>
<td>12.7</td>
<td>The ODL schema definitions for the simple conceptual data model</td>
<td>148</td>
</tr>
<tr>
<td>12.8</td>
<td>Structured type declarations</td>
<td>150</td>
</tr>
<tr>
<td>12.9</td>
<td>Table declarations using structured types and collections</td>
<td>150</td>
</tr>
</tbody>
</table>
Figures and tables

Table D.1 Restricted terms used in the naming of entity types 191
Table D.2 Restricted terms used in the naming of domains 192
Table D.3 Restricted terms used in the naming of attributes 192
Table D.4 Restricted terms used in the naming of relationships 193
Table D.5 Examples of formal attribute names 194
Table E.1 A-priori algorithm: Step 1 results 201
Table E.2 A-priori algorithm: Step 2 results 202
Table E.3 A-priori algorithm: Step 3 results 202
Table E.4 A-priori algorithm: Step 4 results 203
Table E.5 A-priori algorithm: Step 5 results 203
Table E.6 A-priori algorithm: Step 6 results 204
About the author

Keith Gordon was a professional soldier for 38 years, joining the Army straight from school at 16 and retiring on his 55th birthday. During his service he had a number of technical, educational and managerial appointments and gained a Higher National Certificate in Electrical and Electronic Engineering, a Certificate in Education from the University of London Institute of Education, a BA from the Open University and an MSc from Cranfield Institute of Technology. From 1992 until his retirement in 1998, he was first a member of and then head of the Army’s data management team.

He is now an independent consultant and lecturer specialising in data management and business analysis. As well as developing and teaching commercial courses, he is also a tutor for the Open University.

He is a Chartered Member of the British Computer Society and a Member of the Chartered Institute of Personnel and Development.

He holds the Diploma in Business Systems Development specialising in Data Management from the Information Systems Examination Board (ISEB) and he is now a member of their Business Systems Development Examination Panel.

He is the secretary of the Data Management Specialist Group of the British Computer Society and is both a founder member and current committee member of the UK chapter of DAMA International, the worldwide association of data management professionals.
The author of this book is a soldier through and through – but he also has a comprehensive understanding of the principles of data management and is a highly skilled professional educator. This rather unusual blend of experience makes this book very special.

Data management can be seen as a chore best left to people with no imagination but Keith Gordon taught me that it can be a matter of life and death.

We all know that any collective enterprise must have records that are both reasonably accurate and readily accessible. In a commercial operation, failures in data management can lead to bankruptcy. In a public service it can put the lives of thousands of people at risk and waste public money on a grand scale. For a soldier in the heat of battle, any weakness in the availability, quality or timeliness of information can lead to a poor decision that may result in disaster.

So what has this to do with the ‘principles of data management’? It serves as a reminder that a computer application is only as good as the data on which it depends.

It is common for the development of computer systems to start from the desired facilities and work backwards to identify the objects involved and so to the data by which these objects are described. One bad result of this approach is that the data resource gets skewed by the design of specific facilities that it is required to support.

When the business decides that these facilities have to be changed, the data resource must be modified. Does this matter? Some people would say ‘Oh, it’s easy enough to add another column to a table – no problem.’ But these are the same people that get bogged down in the soul-destroying tasks of data fill and the mapping of one database onto another.

There is another way. We don’t have to treat data design as a minor detail understood only by the programmers of a single system. An enterprise can choose to treat its data as a vital corporate asset and take appropriate steps to ensure that it is fit for purpose. To do this it must draw on the body of practical wisdom that has been built up by those large organisations that have already taken this message to heart. The British Army is one such organisation and it was Keith Gordon that made this happen.

The big issue here is how to ensure that the records on which an enterprise depends remain valid and useful beyond the life of individual systems and facilities. This requires good design resting on sound principles validated through extensive practical experience. We live in a changing world where new demands for information are arising all the time. Whether this is due to
new technology, new social problems or the pressures of competition, these new demands cannot be met by creating yet more stove-pipe systems.

The goal we should aim at is for all data to be captured in digital form once only, as close as possible to the time and place of the observations, decisions and results that it is required to reflect. Once captured it should then be stored and distributed in such a manner that it can be made readily available to any person or system with a legitimate ‘need to know’ while remaining safe from loss, damage or theft.

The tricks of the trade through which the best practitioners contrive to bring this about are well documented in this book. I commend it to all people who seek to understand what is involved as well as those who aspire to develop the necessary skills.

Harry Ellis FBCS CITP
Independent consultant and member of W3C
Little Twitchen
Devon, UK
Access control  The ability to manage which users or groups of users have the privilege to create, read, update or delete data that is held in a database.

Attribute  Any detail that serves to qualify, identify, classify, quantify or express the state of a relation or an entity.

Boyce–Codd normal form (BCNF)  In relational data analysis, a relation is in Boyce–Codd normal form if every determinant is a candidate key.

CASE  Acronym for Computer Aided Software Engineering – a combination of software tools that assist computer development staff to engineer and maintain software systems, normally within the framework of a structured method.

Column  The logical structure within a table of a relational database management system (RDBMS) that corresponds to the attribute in the relational model of data.

Conceptual data model  A detailed model that captures the overall structure of organisational data while being independent of any database management system or other implementation consideration – it is normally represented using entities, relationships and attributes with additional business rules and constraints that define how the data is to be used.

Corporate data model  A conceptual data model whose scope extends beyond one application system.

Data  A re-interpretable representation of information in a formalised manner suitable for communication, interpretation or processing.

Data administration  A role in data management concerned with mechanisms for the definition, quality control and accessibility of an organisation’s data.

Data dictionary  Software in which metadata is stored, manipulated and defined – a data dictionary is normally associated with a tool used to support software engineering.

Data management  A corporate service which helps with the provision of information services by controlling or co-ordinating the definitions and usage of reliable and relevant data.
Data mining  The process of finding significant, previously unknown, and potentially valuable knowledge hidden in data.

Data model  (i) An abstract, self-contained logical definition of the data structures and associated operators that make up the abstract machine with which users interact (such as the relational model of data). (ii) A model of the persistent data of some enterprise (such as an entity–relationship model of data required to support the human resources department of Jameson Wholesale Limited – the example used in Chapter 2).

Data modelling  The task of developing a data model that represents the persistent data of some enterprise.

Data owner  (i) The owner of a data definition is the person in the organisation who has the authority to say that this data should be held and that this definition is the appropriate definition for the data. (ii) The owner of a data value is the person or organisation that has authority to change that value.

Data profiling  A set of techniques for searching through data looking for potential errors and anomalies, such as similar data with different spellings, data outside boundaries and missing values.

Data quality  The state of completeness, validity, consistency, timeliness and accuracy that makes data appropriate for a specific use.

Data recovery  Restoring a database to a state that is known to be correct after a failure.

Data security  Protecting the database against unauthorised users.

Data steward  The person who maintains a data definition on behalf of the owner of the data definition.

Data warehouse  A specialised database containing consolidated historical data drawn from a number of existing databases to support strategic decision making.

Database  (i) An organised way of keeping records in a computer system. (ii) A collection of data files under the control of a database management system.

Database administration  A role in data management concerned with the management and control of the software used to access physical data.

Database management system (DBMS)  A software application that is used to create, maintain and provide controlled access to databases.
Datatype  A constraint on a data value that specifies its intrinsic nature, such as numeric, alphanumeric, date.

Discretionary access control (DAC)  Access control where the users who are granted access rights are allowed to propagate those rights to other users.

Domain  A pool of values from which an attribute must take its value – a domain provides a set of business validation rules, format constraints and other properties for one or more attributes that may exist as a list of specific values, as a range of values, as a set of qualifications, or any combination of these.

Enterprise architecture  A process of understanding the different elements that make up the enterprise, such as the people, the information, the processes and the communications, and how those elements interrelate.

Enterprise resource planning (ERP) software  A software package that provides a single integrated database that is planned to meet an organisation’s entire data needs for the management of its resources.

Entity  In a conceptual data model, a named thing of significance about which information needs to be held in support of business operations.

First normal form (1NF)  In relational data analysis, a relation is in first normal form if all the values taken by the attributes of that relation are atomic or scalar values – the attributes are single-valued or, alternatively, there are no repeating groups of attributes.

Foreign key  One or more attributes in a relation (or columns in a table) that implement a many-to-one relationship that the relation (or table) has with another relation (or table) or with itself.

HTML  Acronym for HyperText Markup Language – the markup language used to convey the way that a document is presented by a web browser.

IEC  Acronym for the International Electrotechnical Commission – collaborates with ISO in the development of international standards for information systems.

Information  (i) Something communicated to a person. (ii) Knowledge concerning objects, such as facts, events, things, processes, or ideas, including concepts, that have a particular meaning within a certain context.

Information management  The function of managing information as an enterprise resource, including planning, organising and staffing, and leading, directing, and controlling information.
**Information resource management**  The concept that information is a major corporate resource and must be managed using the same basic principles used to manage other assets.

**Information system**  A collection of manual and automated components that manages a specific information resource.

**ISO**  Acronym for the International Organization for Standardization – collaborates with IEC in the development of international standards for information systems.

**Mandatory access control (MAC)**  Access control where access rights cannot be changed by the users.

**Metadata**  Data about data – that is, data describing the structure, content or use of some other data.

**Multi-level security**  The ability of a computer system to process information with different security levels, to permit access by users with different security clearances, and to prevent users from obtaining access to information for which they do not have authorised access.

**Multimedia data**  Data representing documents, audio (sound), still images (pictures) and moving images (video).

**Normal form**  A state of a relation that can be determined by applying simple rules regarding dependencies to that relation.

**Normalisation**  Another name for relational data analysis.

**Object orientation**  A software-development strategy based on the concept that systems should be built from a collection of reusable components called objects that encompass both data and functionality.

**ODMG**  Acronym for the Object Data Management Group, a body that has produced a specification for object-oriented databases.

**OLAP**  Acronym for online analytical processing – a set of techniques that can be applied to data to support strategic decision making.

**OLTP**  Acronym for online transactional processing – data processing that supports operational procedures.

**Primary key**  The set of mandatory attributes in a relation (or mandatory columns in a table) that is used to enforce uniqueness of tuples (or rows).

**RDBMS**  Acronym for relational database management system – a database management system whose logical constructs are derived from the relational model of data. Most relational database management systems
available are based on the SQL database language and have the table as their principal logical construct.

**Relation**  The basic structure in the relational model of data – formally a set of tuples, but informally visualised as a table with rows and columns.

**Relational data analysis**  A technique of transforming complex data structures into simple, stable data structures that obey the rules of relational data design, leading to increased flexibility and reduced data duplication and redundancy – also known as normalisation.

**Relational model of data**  A model of data that has the relation as its main logical construct.

**Relationship**  In a conceptual data model, an association between two entities, or between one entity and itself.

**Repository**  Software in which metadata is stored, manipulated and defined – a repository is normally associated with a corporate data management initiative.

**Repository administration**  A role in data management concerned with the management and control of the software in which ‘information about information’ is stored, manipulated and defined.

**Schema**  A description of the overall logical structure of a database expressed in a data definition language (such as the data definition component of SQL).

**Second normal form (2NF)**  In relational data analysis, a relation is in second normal form if it is in first normal form and every non-key attribute is fully functionally dependent on the primary key – there are no part-key dependencies.

**SQL**  Originally, SQL stood for structured query language. Now, the letters SQL have no meaning attributed to them. SQL is the database language defined in the ISO/IEC 9075 set of international standards, the latest edition of which was published in 2003. The language contains the constructs necessary for data definition, data querying and data manipulation. Most vendors of relational database management systems use a version of SQL that approximates to that specified in the standards.

**Structured data**  Data that has enforced composition to specified datatypes and relationships and is managed by technology that allows for querying and reporting.

**Table**  The logical structure used by a relational database management system (RDBMS) that corresponds to the relation in the relational model of data – the table is the main structure in SQL.
**Third normal form (3NF)**  In relational data analysis, a relation is in third normal form if it is in second normal form and no transitive dependencies exist.

**Tuple**  In the relational model of data, the construct that is equivalent to a row in a table – it contains all the attribute values for each instance represented by the relation.

**Unified Modeling Language (UML)**  A set of diagramming notations for systems analysis and design based on object-oriented concepts.

**Unstructured data**  Computerised information which does not have a data structure that is easily readable by a machine, including audio, video and unstructured text such as the body of a word-processed document – effectively this is the same as multimedia data.

**XML**  Acronym for eXtensible Markup Language – the markup language used to convey the definition, structure and meaning of the information contained in a document.
Preface

I think I first decided that I wanted to be a soldier when I was about three years of age. In 1960, aged 16 and with a slack handful of GCE ‘O’ Levels, I joined the Royal Armoured Corps as a junior soldier. I suppose I thought that driving tanks would be fun, but my time with the Royal Armoured Corps was short-lived and, in 1962, I joined the Royal Corps of Signals and trained as an electronics technician. I learned to repair and maintain a range of electronics equipment that used logic AND, OR, NAND and NOR gates, multivibrators, registers and MOD-2 adders, all of which are the building blocks of the central processing units at the heart of computers. Nine years later, I attended a course that turned me into a technical supervisor. This course extended my knowledge to include the whole range of telecommunications equipment. I now knew about radio and telephony as well as being the proud owner of a Higher National Certificate in Electrical and Electronic Engineering. On this course we also met a computer, an early Elliot mainframe, and learned to program it. After this course I found myself in Germany with a brilliant job, responsible for the ‘system engineering’ of the communications for an armoured brigade headquarters. Not only was I ensuring that my technicians kept the equipment on the road, but I was also designing and having my staff build the internal communications of the headquarters – which involved the interconnection of about a dozen vehicles.

A career change happened in 1978 when, following a year’s teacher training, I was commissioned into the Royal Army Educational Corps. I spent the next nine years in classrooms in Aberdeen, London, the Falkland Islands (not sure that some of the places where I taught when there could be called classrooms, but...) and Beaconsfield. In Beaconsfield, I taught maths, electronics and science; in the other jobs, I taught a mixture of literacy, numeracy, current affairs and management. It was these teaching jobs that gave me my greatest sense of personal satisfaction. I also extended my knowledge of computing by studying for a BA with the Open University. 1987 saw me getting deeper into computing by studying for an MSc in the Design of Information Systems, where I was introduced to databases and structured methods. I left the course thinking I knew about data and data modelling. I now know that I had hardly scraped the surface.

In 1992, after two more educational jobs, I was offered a job in ‘data management’. Well, I knew about ‘data’ and I had taught ‘management’ so, despite never having before heard the two words used together, I thought it sounded like my thing. I may have been influenced by the belief that the job would involve an office in London which was close enough to home to commute daily. It came as shock to find that the office was in Blandford, where I had
already served for over seven years during my time in the Signals, and it severely disrupted my home life. But this was nothing unusual; disruption of home life is a substantial part of the lot of a soldier.

The Army had commissioned one of the large consultancy companies to conduct a major study into its information systems. This study had recommended that the Army should have a data management team and this team came together in 1992. There were five of us: four officers and a civil servant. All we knew was that data management was to be good for the Army. Nowhere was there a description of what data management was. So we were in a highly desirable position: we had to work out what we had to do. I think this period provided me with the greatest technical challenge of my Army career. What I was aware of was that the Army had a large number of information systems, all independently designed, and it was virtually impossible to share information between them. And the Army was also undertaking a large programme of information systems procurement, in some important cases into areas that had not previously had information systems support. To make the Army more effective on the battlefield and, at the same time, to reduce our casualties, it was vital that the information systems could share information. The Army had a vision of a single, fully integrated information system. This would not, of course, be a single system, but a federation of systems that appeared to the user as a single system. This could not be achieved without data management.

Thus began my interest in data management. Three years later I was promoted and became the head of the team until I retired from the Army in 1998. I now work as an independent consultant and lecturer. As well as teaching commercial courses in data management and business analysis, I have also been a tutor with the Open University since 1999, tutoring database and general computing courses in the undergraduate and postgraduate programmes. My data management journey continues.

I believe that all medium to large organisations, commercial and government, need a corporate data management service. I see many instances where the inability to share information between information systems leads to mistakes and misunderstanding, which in turn leads to poor customer service (even government departments have customers) and extra expenditure. These organisations cannot really afford to be without data management, yet very few recognise the problems, let alone that data management is the solution. Regrettably, this ignorance exists not only amongst business managers; it is rare to find an IT or IS manager who sees the need for data management. In fact most, like me 14 years ago, have never heard the two words ‘data’ and ‘management’ used together. I hope that this book goes some way to bring data management to the attention of those who really ought to know about it.

This book, therefore, represents the knowledge I have gained over the last 14 years. Some of this knowledge came from doing the job, some from the
people I have taught and some from the many books sitting on my bookshelves, most of which are listed either as specific references or as suggestions for further reading.

I owe a debt of gratitude to a number of people who have helped me on my data management journey. Ian Nielsen, Martin Richley, Duncan Broad and Tim Scarlett were my colleagues in that original Army data management team who shared those many hours around a whiteboard trying to work out what it was all about. There were others involved as well. David Gradwell and Ken Allen were our first consultants, introducing us to the mysteries of metadata models and naming standards. Later on when we started data modelling in earnest we had the benefit of the experience of Harry Ellis and Ron Segal (who is now in New Zealand). I learnt masses from working with all of these people and I think we were all (including our experienced consultants) on a learning curve. At the start of my data management journey, I attended a Principles of Data Administration course run by Chris Newton for Stehle Associates. This course set data management and data administration in context. It is Chris’s Principles of Data Administration course that is the skeleton on which I have built my own Principles of Data Management course which I now deliver for Stehle Associates. Dave Beaumont, the principal of Stehle Associates, has encouraged me to develop data management courses and he and I have spent many hours discussing data management issues. He has also kindly reviewed early drafts of some of the chapters of this book. Thanks too to Ian Sinclair, one of my colleagues on the committee of the UK chapter of DAMA International, who reviewed the chapter on data quality; to Matthew West, who reviewed the appendix on generic data models; and to Tony Jenkins who has reviewed the whole book and provided many useful recommendations for its improvement. I would also like to thank the many people I have not mentioned but whom I have either worked with or discussed data management issues with over the last 14 years. I have learnt from you all.

Particular thanks are due to Matthew Flynn, Suzanna Marsh and Florence Leroy of the British Computer Society who have been instrumental in getting this book into print.

Finally, a massive thank you to my wife, Vivienne, for her unstinting support over the last 40 years. Being a soldier’s wife for 32 years was never going to be a picnic and she had a right to expect things to be more relaxed and easier when I retired. Instead, with consulting, teaching and, now, the writing of this book, I have neglected her far more than I should have done and still she is there looking after me.

Keith Gordon
High Wycombe
January 2007
Introduction

This book is called *Principles of Data Management* but it is really about having the policies and procedures in place within an organisation so that the various information systems that the organisation uses to support its activities can provide high-quality information to their users, even if that information did not originate in the information system with which the user is currently interacting. For this to happen, the organisation's information systems must be able to share information. If there is no automatic sharing of information between the information systems, some departments may be kept in the dark about what is going on in other departments and information may have to be keyed into more than one system. Neither of these situations helps the organisation’s effectiveness or efficiency.

The key to the provision of high-quality information and the sharing of information between information systems is to have an effective corporate data management policy in place. Yet very few senior business and IT or IS managers have heard of data management, let alone have an effective data management policy in place.

This book is aimed at three audiences. First, there are the data management practitioners themselves. They are presumably already committed to data management but may be struggling to find all the information that they need to set their role in the wider business context or to perform the myriad tasks that fall within the scope of data management.

This book will not have all the answers, but it may provide an indication of what the answer should be and, perhaps, where to go and look for the answer. Secondly, there are the IT or IS managers who have heard of data management, are probably aware that it might be a good idea, but are not sure what it involves or what the implications of having a corporate data management function will be. Maybe they already have a data management team working within their department but are not sure what that team does or what it should do. The third group who should read this book – or, at least, the sections that are not too technical – are the business managers who want to understand why they are being asked to pay for a team of data managers who do not look as if they are going to deliver the much-sought-after return on investment within the current budgetary cycle.

For the data management practitioners, I commend the data management qualifications offered by the Information Systems Examination Board of the British Computer Society. At the time of writing, two qualifications are provided, a Certificate in Data Management Essentials and a Diploma in Data Management, but it is anticipated that additional certificates will
become available. The certificate level is examined by a short written examination and the diploma level is examined by an oral examination. This book covers the existing syllabus for the Certificate in Data Management Essentials as well as providing additional material for anyone proposing to take the Diploma in Data Management examination.

So, to meet the requirements of practitioners, IT or IS managers and business managers, this book covers the whole range of data management activities. There are 12 chapters and eight appendices:

- Chapter 1 – Data and the Enterprise – introduces the idea that information is a key business resource. It starts by exploring the relationship between information and data. We then move on to a discussion of the importance of the quality of the data that underlies the information. If the quality of data is important what are the common problems with data? Why must we take an enterprise-wide view of data? The chapter concludes by highlighting that the management of data is a business issue and not a technical issue.

- Chapter 2 – Database Development – is a long, largely technical, chapter that provides an explanation of how the databases at the heart of all information systems are designed. It introduces the concepts of database architecture and then provides examples of two analysis techniques – conceptual data modelling and relational data analysis – and how these lead to a physical database design.

- Chapter 3 – What is Data Management? – first considers the problems encountered without data management then introduces the scope of the responsibilities of data management. We then look at the three separate roles within data management – data administration, database administration and repository administration. We end this chapter by summarising the benefits of data management.

- Chapter 4 – Corporate Data Modelling – looks at data modelling when applied to an enterprise’s total data requirements as opposed to being applied to the smaller set of requirements that are to be met by a single information system. We explain why corporate data models are required and then introduce some more data modelling concepts. We discuss where corporate data models should lie on the continuum from abstract to detailed. We then suggest how the development of a corporate data model may be approached and introduce six principles to be applied to the development of corporate data models.

- Chapter 5 – Data Definition and Naming Conventions – introduces the key data definition and naming ‘standards’ used by data managers. We discuss the principles underlying these standards and provide some examples.
• Chapter 6 – Metadata – introduces the concept of ‘data about data’ and the way that it is used.

• Chapter 7 – Data Quality – provides an overview of this important area. We define the term ‘data quality’, we look at how poor-quality data can affect a business, we consider what causes poor-quality data and we look at techniques for improving data quality. The fact that the achievement of data quality requires an ongoing procedural and cultural change, and not just a one-off project, is stressed.

• Chapter 8 – Data Accessibility – brings together in one chapter the related issues of data security, protecting the database against unauthorised users, data integrity, protecting the database against authorised users, and data recovery, bringing the database to a usable consistent state after a failure.

• Chapter 9 – Database Administration – provides an overview of the roles and responsibilities of database administrators, particularly the monitoring and tuning of the performance of a database.

• Chapter 10 – Repository Administration – looks at the management and control of the software in which ‘information about information’ is stored, manipulated and defined.

• Chapter 11 – The Management of Data Management – describes the knowledge and skills required for each of the three data management roles: data administration, database administration and repository administration. We then discuss where in the organisational hierarchy the data management function and its subordinate elements should be placed.

• Chapter 12 – Industry Trends and their Effects on Data Management – covers a number of fads, advances and developments, including the recent developments in SQL. Data management practitioners should not only be aware of these trends, but should ensure that their organisations have policies in place to take account of these developments. The trends considered are:
  ✷ the use of software application packages, such as accounting packages;
  ✷ distributed data and databases;
  ✷ data warehousing and data mining;
  ✷ object orientation and databases;
Introduction

- multimedia databases;
- data and web technology.

- Appendix A – Comparison of Data Modelling Notations – looks at some alternatives to the data modelling notation used throughout the book.

- Appendix B – Hierarchical and Network Databases – looks at two popular pre-relational database models and their implementations.

- Appendix C – Generic Data Models – looks at why data models become generic (or abstract) and the advantages and disadvantages of using generic data models as the basis for database design.

- Appendix D – An Example of a Data Naming Convention – provides a complete example of a data naming convention.

- Appendix E – Metadata Models – looks at the data models that underpin data dictionaries and repositories.

- Appendix F – A Data Mining Example – provides a worked example of just one of the many data-mining techniques that are available.

- Appendix G – HTML and XML – looks in more detail at these two key ‘technologies’ used with web technology.

- Appendix H – XML and Relational Databases – looks at the support for XML provided in the SQL standard.
Data and the Enterprise

This chapter introduces the concepts of information and data and discusses why they are important business resources within the enterprise. The problems caused by data which is of poor quality or is inconsistent are introduced.

INFORMATION IS A KEY BUSINESS RESOURCE

When asked to identify the key resources in any business, most business people will readily name money, people, buildings and equipment. This is because these are the resources that senior business managers spend most time managing. This means that in most businesses there is a clear investment by the business in the management of these resources. The fact that these resources are easy to manage and that the management processes applied to these resources can be readily understood by the layman means that it is seen to be worthwhile investing in their management. It is usually easy to assess how much the business spends on managing these resources and the return that is expected from that investment.

But there is a key resource missing from that list. That missing resource is ‘information’. Without information, the business cannot function. Indeed, the only resource that is readily available to senior management is information. All important decisions made within an enterprise are based on the information that is available to the managers.

Despite its importance, most business people do not recognise information as a key business resource. Because of its association with technology (with ‘information technology’ having become in effect one word, generally with more emphasis on the ‘technology’ than on the ‘information’), information is seen as something mystical that is managed on behalf of the business by the specialist Information Technology or Information Systems department. The management of information is seen, therefore, as something requiring special skills beyond the grasp of the layman. It is very difficult to determine how much the business spends on managing information or, indeed, the return it can expect from that expenditure.

Information is a business resource that is used in every aspect of a business: it supports the day-to-day operational tasks and activities; it enables the routine administration and management of the business; and it supports strategic decision making and future planning.

For a supermarket chain, the operational tasks and activities include the processing of customers’ purchases through the electronic point-of-sale system and the ordering of goods from suppliers; for a high street bank, they
include the handling of customers’ cash and cheques by the cashiers, the processing of transactions through ATMs and the assessment of the credit status of a customer who is requesting a loan; for an online book ‘store’, they include the collection of customers’ orders, the selection and dispatch of the books and the production of a customer profile enabling the ‘store’ to make recommendations to customers as they log on to the website.

For all types of business, information in various forms is routinely used by managers to monitor the efficiency and effectiveness of the business. Some of this information comes in the form of standard reports. Other information may come to the managers as a result of their ad-hoc questions, maybe directed to their subordinates but, increasingly, directed to the information systems that support the business.

All businesses need to plan for their future and take high-level, strategic decisions. In some cases the consequence of making an incorrect strategic decision could be the ultimate collapse of the business. To carry out this future planning and strategic decision making, the senior management of the business relies on information about the historic performance of the business, the projected future performance of the business (and this, to a large extent will be based on an extrapolation of the historic information into the future), their customers’ present and future needs, and the performance of their competitors. Information relating to the external environment, particularly the economy, is also important. For a supermarket chain, these decisions may include whether to diversify into, say, clothing; for a high street bank, they may include the closure of a large number of branches; and for an online book ‘store’ whether to open new operations overseas.

Information is important, therefore, at every level in the business. It is important that the information is managed and presented in a consistent, accurate, timely and easily understood way.

THE RELATIONSHIP BETWEEN INFORMATION AND DATA

Wisdom, knowledge, information and data are all closely related through being on the same continuum – from wisdom, to knowledge, then to information and, finally, to data. This book is about managing data to provide useful information, so we will concentrate on the relationship between information and data.

An often-heard definition of information is that it is ‘data placed in context’. This implies that some information is the result of the translation of some data using some processing activity, and some communication protocol, into an agreed format that is identifiable to the user. In other words, if data has some meaning attributed to it, it becomes information.

For example, what do the figures ‘190267’ represent? Presented as ‘19/02/67’, it would probably make sense to assume that they represent a date. Presented on a screen with other details of an employee of a company, such as name and address, in a field that is labelled ‘Date of Birth’ the meaning becomes
obvious. Similarly, presented as ‘190267 metres’, it immediately becomes obvious that this is a long distance between two places but, for this to really make sense, the start point and the end point have to be specified as well as, maybe, a number of intermediate points specifying the route.

Whilst these examples demonstrate the relationship between data and information, they do not provide a clear definition of either data or information.

There are many definitions of data available in dictionaries and textbooks but, in essence, most of these definitions basically say that data is ‘facts, events, transactions and similar that have been recorded’. Furthermore, as I pointed out earlier, the definition of information is usually based on this definition of data. Information is seen as data in context or data that has been processed and communicated so that it can be used by its recipient.

The idea that data is a set of recorded facts is found in many books on computing. However, this concept of data as recorded facts is used beyond the computing and information systems communities. It is, for example, also the concept used by statisticians. Indeed, the definition of data given in Webster’s 1828 Dictionary – published well before the introduction of computers – is:

**Things given, or admitted; quantities, principles or facts given, known, or admitted, by which to find things or results unknown.**

However, starting the development of our definitions by looking at data first appears to be starting at the wrong point. It is information that is important to the business, and it is there that our definitions, and our discussion about the relationship between information and data, should really start.

We start by considering the everyday usage of information – something communicated to a person – and, with that, we can have a definition of data that is relevant to the theme of this book. Data is ‘a re-interpretation of information in a formalised manner suitable for communication, interpretation or processing’ (ISO/IEC 2382-1, 1993). There is a note attached to this definition in the ISO/IEC standard which states that data can be processed by human or automatic means; so this definition covers all forms of data but, importantly, includes data held in information systems used to support the activities of an organisation at all levels: operational, managerial and strategic.

Figure 1.1 provides an overview of the relationship between data and information in the context of a computerised information system. The user of the system extracts the required information from their overall knowledge and inputs the information into the system. As it enters the system, it is converted into data so that it can be stored and processed. When another system user requires that information, the data is interpreted – that is, it has meaning applied to it – so that can be of use to the user.
For most of this book, we consider data stored in a database. This is often called ‘structured data’. However, it must be understood that a considerable proportion of an organisation’s information may be held in information systems as ‘unstructured data’ – in word-processed documents, drawings, etc.

THE IMPORTANCE OF THE QUALITY OF DATA

Since information is an important resource for any organisation, information presented to users must be of high quality. The information must be up to date, complete, sufficiently accurate for the purpose it is required, unambiguously understood, consistent and available when it is required.

It is essential that information is up to date. When customers buy their shopping at the supermarket they need to be charged the current price for the items they have bought, not the price that was current yesterday before the start of today’s cut-price promotion. Similarly, managers reordering stock need to be aware of the current, not last week’s, stock levels in order to ensure that they are not over or under-stocked.

Only when the information available is complete can appropriate decisions be made. When a bank is considering a request for a loan from a customer, it is important that full details of the customer’s financial position is known to safeguard both the bank’s and the customer’s interests.

Information on which important decisions are made must be accurate; any errors in the potential loan customer’s financial information could lead...
to losses for the bank, for example. Whilst it is important that information is accurate, it is possible for the information to be ‘too accurate’, leading to the information being misinterpreted. Earlier I quoted ‘190267 metres’ as the distance between two points, say London and Birmingham. But the figure ‘190267’ implies that this distance has been measured to the nearest metre. Is this realistic? Would it not be more appropriate to quote this figure as ‘190 kilometres (to the nearest 10 kilometres)’? I cannot answer that question without knowing why I need to know the distance between London and Birmingham. Information should be accurate, but only sufficiently accurate for the purpose for which it is required.

To be accurate from a user perspective, information must also be unambiguously understood. There should be no doubt as to whether the distance the user is being given is the straight-line distance or the distance by road. The data should also be consistent. A query asking for the distance between London and Birmingham via a specified route should always come up with the same answer.

Information has to be readily available when and where it is required to be used. When it is time to reorder stock for the supermarket then the information required to decide the amount of replacement stock to be ordered has to be available on the desk of the manager making those decisions.

Information is derived from the processing of data. It is vital, therefore, that the data we process to provide the information is of good quality. Only with good-quality data can we guarantee the quality of the information. Good-quality data is data that is accurate, correct, consistent, complete and up to date. The meaning of the data must also be unambiguous.

THE COMMON PROBLEMS WITH DATA

Unfortunately, in many organisations there are some major, yet unrecognised or misunderstood, data problems. These problems are generally caused by a combination of the proliferation of duplicate, and often inconsistent, occurrences of data and the misinterpretation and misunderstanding of the data caused by the lack of a cohesive, enterprise-wide data definition regime.

Whenever it is possible for any item of information to be held as data more than once, there is a possibility of inconsistency. For example, if the addresses of customers are held in more than one place – or, more specifically, in more than one information system – and a customer informs the company that they have changed their address, there is always the danger that only one instance of the address is amended, leaving the other instances showing the old, incorrect address for that customer. This is quite a common scenario. The marketing department and the finance department may have separate information systems: the marketing department has a system to help it track customers and potential customers whilst the finance department has a completely separate system to support its invoicing and payments received accounting functions. With information systems independently designed
and developed to support individual business areas or specific business processes, the duplication of data, and the consequent likelihood of inconsistency, is commonplace. Unfortunately, in most organisations, the potential for inconsistency through the duplication of data is getting worse because of the move away from centralised mainframe systems, the proliferation of separate departmental information systems and the availability of personal desktop computing power, including the provision of spreadsheet and database software.

Even where it is understood that it would be to the advantage of the organisation for information to be shared between these separate systems, this is often impossible without there being the possibility of misinterpretation or misunderstanding of the information that is shared.

In its 1994 publication ‘Corporate Data Modelling’, the Central Computer and Telecommunications Agency – now part of the Office of Government and Commerce – recognised that there are a number of possible reasons for sharing information. These are:

- when central reference data is used by independent operational units, such as product codes and product prices;
- when public domain datatypes are used and exchanged, for example, when publicly available statistical data sets are to be used;
- when operational results need to be collated across several profit centres, for example, to collate or compare the sales figures from stores within a supermarket chain;
- when the output from one system forms the input to another, for example, the output of a forecasting system is used by another system to determine resource and budget implications;
- when application systems performing a similar function for distinct autonomous units are required to harmonise their data to permit close collaboration, for example, the command and control systems for the police, fire and ambulance services need to ‘work together’ in the event of an emergency.

The sharing of information between independently designed and developed information systems is technically straightforward. It is a relatively simple matter to electronically connect two or more information systems together using a network and then to transfer data between them. The difficulties come after the data has been transferred and the receiving information system cannot interpret the data or, worse still, interprets the received data according to its understanding of the meaning of the data, but this interpretation differs from that used in the originating system. This possibility of the misinterpretation of transferred data is very common in organisations and the situation is getting worse.

This is also a consequence of the proliferation of independently designed and developed departmental or single-function information systems. At the heart of an information system is a database whose purpose is to provide
persistent storage of the data. Each of these databases is designed to ensure that the data is available when required by the applications supported by that information system and, possibly, to maintain the integrity of the data within that particular database. A database is designed to provide effective and efficient support to the business area or function that the information system is being designed to support by meeting the immediate data requirements for that business area or function as they are understood by the database designer. It is very rare for a wider view of current or future data requirements to be taken.

The proliferation of departmental or function-specific information systems, each with its own database designed without recognition of wider data requirements, has led to widespread problems of data: inconsistency caused by duplication across different information systems and misinterpretation when data is shared between information systems.

AN ENTERPRISE-WIDE VIEW OF DATA

In order to improve the quality of information across an organisation, we must first understand the data that provides that information and the problems that are associated with that data. We must also look at business information needs and move the organisation to a position where the required data is made available to support the current information needs in a cost-effective manner whilst providing the flexibility to cope with future needs in a reasonable time scale. We need to consider the information needs of the whole organisation and then manage the data in such a way that it supports the organisation’s total information needs.

In order to manage the organisation’s data resources effectively we must first understand it. This requires more than just recognising data as being the raw material in the production of information. It implies knowledge of what data is important to the business, and where and how it is used. What functions and processes use the data? When is it created, processed and destroyed? Who is responsible for that data in all stages of its life?

It is also essential that we produce a clear and unambiguous definition of all data that the organisation uses. Such a definition must be a common view, accepted and agreed by all business areas.

Effective management of data also requires an understanding of the problems that relate to data. These problems often cross departmental boundaries and their solutions consist of both technical and organisational aspects.

Organisations vary tremendously in size and nature. A large multinational organisation tends to have different data-related problems from a small company although, even in a small company, the problems can be quite complex. The type of business may also affect the nature of the problems. A large proportion of the information systems in a finance or insurance company relate to customers or potential customers. In a manufacturing environment, however, dealing with customers is only one part of the overall business processes.
At the more technical level, data-related problems are affected by the types of computer system in place. Are the systems networked or distributed? Is extensive use made of personal computers? Are there multiple computer sites? And so on.

Individual departments do not necessarily perceive a given problem as having a potential impact across the whole organisation. One of the difficulties often faced by a central team responsible for managing the data for the whole organisation is bridging the gap between different departmental views. This requires patience and tact. It certainly requires authority, or access to appropriate authority, as the implementation of a solution may well involve co-operation with several managers within the organisation. Most importantly, it demands an understanding both of the information needs of the whole business and of the nature of the associated technical and organisational problems.

In reality, the problems relating to data are often very complex and affect many different areas within an organisation. Data is used in different ways by different business functions. Data can take many forms and the technologies for handling and storing data are constantly changing. Data problems do not appear in a form that enables a neatly packaged, stand-alone solution for the handling and management of data.

Recently a number of vendors have been supplying enterprise resource planning (ERP) software, which is supposed to provide a single integrated database that meets an organisation’s entire data needs for the management of its resources. In general, these products do not appear to be providing the advantages claimed. Unless the organisation is prepared to replace all of its information systems in one go there will still be a need for the data held by the enterprise resource planning system to be integrated with the data held by the existing information systems that are still in use. Also, to really take advantage of enterprise resource planning software the organisation probably needs to change its business processes to conform to the processes supported by the software, and many businesses are not prepared to make these changes.

**MANAGING DATA IS A BUSINESS ISSUE**

We identified money, people, buildings and equipment as the key resources in any business and we added information to that list.

For all of these resources some special responsibilities exist within the organisation:

- The finance department has special responsibilities for managing the organisation’s money, including the allocation of budgets, managing investments and accounting.
- The personnel department has special responsibilities for managing the organisation’s employee base, including the provision of advice on legislation affecting personnel issues and the recruitment of staff.
• The estates department has special responsibilities for managing the buildings used by the organisation, including ensuring that the buildings meet legal requirements in respect of health and safety and discrimination issues, buying, selling and leasing of buildings and ensuring that the estate is adequately insured.

• The stores and maintenance department has special responsibilities for managing the organisation’s equipment, including the provision of a central purchasing function, the accounting for equipment in use and the storage of equipment until it is required for use.

• The IT or IS department has special responsibility for data and information, including the physical storage, distribution, security, backup and archiving of data.

In most organisations, it is now common practice for line management to have responsibility for the day-to-day administration and management of these resources, with the specialist departments only providing specialist advice to the line management. People have to be managed on a day-to-day basis; money is allocated to budget holders to use and manage according to specific rules; buildings are run and administered; equipment is used and maintained.

Additionally, information is collected, validated and used. This is very much the responsibility of the business. All the decisions about what is collected and how it is validated are business decisions. So are the decisions about how information is to be handled and stored as data. Any data management function must, therefore, support the business. Data management is not purely a technical issue; the definition of the data to be stored should be the responsibility of the business. Most organisations are counting the cost of ineffective data management. Real business opportunities may be lost as a result of the inability to respond quickly to changing requirements. There are many situations where information exists but is not accessible in the right time frame.

In many cases the only way that information may be shared between information systems is by reading information from one screen and keying it into another system or, worse still, systems. The cost of continually rekeying information in this way is significant both in terms of the resource required to carry out this task and in potential errors through misinterpretation of the information that is to be rekeyed. Such costs impact on the business as well as on the IT or IS department, although the greater impact is on the business. Surprisingly, this approach to information sharing is still in use in some organisations in 2007.

There are many claimed benefits for having a data management function within the organisation. These benefits nearly all make sound business sense and can be recognised as such. However, not all of them can be related to direct cost savings. Consequently, it requires a degree of faith on the part of management that the end result, the benefits, will justify the costs.
The benefits split into two areas: those which are business-oriented and those which are systems-oriented. The former include cost savings through, for example, the reduction in duplicated marketing mailings and improved customer service, whilst the latter include reduced time to develop new applications, which also translates into financial savings. I firmly believe, however, that the systems-oriented benefits are a natural by-product of a business-oriented data management initiative. The reverse is not necessarily true. There may be no additional benefits to business effectiveness and efficiency if the IT or IS function implements data management in order to save on development costs.

It is relatively easy to quantify the costs of today’s problems, both in financial terms and as lost business opportunities. Thus, it is possible to demonstrate relatively easily the potential benefits of reducing or even eradicating such problems and enabling the business to exploit the huge investment it has already made in data for optimum returns. It is possible to make the business case for the establishment of a data management function.

**SUMMARY**

In this chapter we have seen that information, an often neglected key business resource that needs to be shared across an enterprise, is developed from data. To provide quality information, data has to be properly managed. There has to be an enterprise-wide view of data, and the business, not the IT or IS function, has to take the lead in the management of data.
Index

a-priori algorithm 143, 201–204
abbreviations 88, 190–191
access control
authentication procedures 102
definition of xix
discretionary (DAC) xxii, 105
for repositories 121
in SQL 102–105
mandatory (MAC) xxii, 105–106
purpose of 102
ACID properties 109
aggregation 146
Agricultural Metadata Element Set (AgMES) 92
aliases 85
alternate keys 47
association rules 201
asynchronous replication 138
attributes
definition of xix
in conceptual data models 22–23
in relational data models 34
audit trails 102, 107
authentication procedures 102
backups 110
Barker, R. 19, 159
behaviour, of object 145
Bell-LaPadula Model 105
Bell, D. 105
BINARY LARGE OBJECT (BLOB) datatype 154
BMP format, metadata in 93
bottom-up integration 133, 137–138
Boyce-Codd-normal form (BCNF) xix, 47–48
Chen, P. 161
class, of object 145
class terms 88, 193
problems with restriction of 89–90
cluster analysis 143
Codd, E. 33, 169
collection types, in SQL 149
columns
definition of xix
generally 14, 19
naming of 51, 190
Common Gateway Interface (CGI) 157
completeness 97, 99
Computer Aided Software Engineering (CASE) tools xix, 117–118
conceptual completeness 99
conceptual correctness 99
conceptual data models
and database physical design 49–50
and relational data analysis 48
concepts 20–33, 66–72
definition of xix
description of 18
entity supertypes 67
naming standard for 184–187
roles of 49–50
conceptual organisation of data 14
constraints
and entity subtypes 69
and exclusivity 70
declaring 19, 20
integrity 107–108
key 108
containment 147
controlled vocabulary 90
corporate data models
and single project data models 99
definition of xix
development of 74–77
generally 65–66
nature of 72–74
principles of 78
correctness 97, 99
C++ 147
cubes of data
operations on 141–143
structure of 140–141
data
and information 2, 3–4
conceptual organisation of 14
definition of xix
development of 74–77
generally 65–66
nature of 72–74
principles of 78
granularity of 144
problems with 5–8
data administration
definition of xix, 61
position in organisation 125–130
skills 121–125
tasks of 86
database administration
definition of xx, 61
position in organisation 125–130
responsibilities of 112–114
database performance 114–115
database physical design 112–113
education/training 113–114
skills 123–125
database development process
database physical design 19–20
description of 16–18
information requirement analysis 17–19
database logs 110–111, 115
database management systems (DBMS)
definition of xx
functions of 13–14, 51
generally 11, 12, 13
management of 113
security features 102, 106
database object privileges 103
database physical design
conceptual data models and 19–20, 49–50
first-cut design 50–52
optimised design 52–53
role of database administrator 112–113
databases
advantages of 12–13
and web technology 155–157
definition of xx
description of 11, 12
design for quality 96
types of failure 109–110
data cache 115
data clustering 52, 115
data control language (DCL) 102–104
data definition language (DDL) 102, 147
data definitions
development of 58
elements of 84–88
aliases 86
format 84
names 84
ownership xx, 85
significance statements 84
stewardship xix, 86
synonyms 85
validation criteria 84
valid operations 84
valid values 84
in database systems 11
data dictionaries
and CASE tools 116
definition of xix, 195
metadata models for 195
data independence
logical 12, 16
physical 12, 16
data integrity
and quality 108
entity integrity 108
referential integrity 108
using constraints 107–108
DATALINK datatype 154
data management
activities/deliverables 57–59
as business issue 8–9, 54–55, 59
as information service 58
benefits of 9, 62
definition of xix, 54
education about 58
effects of not having 55–57
policy/strategy xix, 58
position within organisation 125–130
qualifications xxix–xxx
relationship with information management 60
responsibilities of 57
Index

roles within 60–62
support for system development 59
data manipulation language (DML) 102–104
data marts 140
data mining
  a-priori algorithm 143, 201–204
  cluster analysis 143
  definition of xx, 143
  neural networks 143
  statistical techniques 143
data models
  definition of xx
  development of 19, 58
  multidimensional 140–143
  quality of 98–99, 178
data naming
  abbreviations in 88, 190–191
  class terms 88–89, 193
  in data definition 84–85
  modifier terms 89, 193
  prime terms 89, 193
  problems with 89–90
  purpose of 88
  standard for 183–194
  thesaurus 90
data ownership xx, 85
data profiling xx, 100
data quality
  and constraints 108
  as business issue 100
  causes of poor quality data 96–97
  data model quality 98–99, 178
  data profiling xx, 100
  definition of xx, 95
  dimensions of 97–98
  importance of 4–5
  improving 99–101
  in repositories 121
  issues with poor quality data 95–96
  Total Quality Data Management (TQdM) 100
data recovery xx, 109–111
data security
  audit trails 102, 107
  authentication procedures 102
  Bell-LaPadula Model 105
  definition of xx, 102
  encryption 107
  multi-level security xxii, 106
  policy 102
data stewardship xx, 86
datatypes
  definition of xx
  generally 50–52
  structured types as 149
  data warehouses
  and OLAP 142–143
  architecture 139–140
  definition of xx
  description of 139
  multidimensional model of data 140–143
  relational schemas for 144–145
  Date, C. 134
  deadlocks 115
denormalisation 53, 115
dependent entities 161–162
dice operation 142
direct access 11
directories 116
discretionary access control (DAC) xxi, 105
disk mirroring 111
distributed databases
  availability 133
  bottom-up integration 133, 137–138
  fragmentation/partitioning 135–137
  fundamental principle of 134
  objectives for 134–135
  rationale for 132
  reliability 133
  replication management 138
  schemas 133–138
  top-down distribution 133, 135–137
Document Type Definition (DTD) 210
DoDAF 63
domains
  and datatypes 51–52
  definition of xxi, 34
  drill-down (de-aggregation) operation 142
  Dublin Core 92
  duplication of data 5–6
education/training
  for database administrators 113–114
  for repository users 121
Ellis, H. 159
enterprise architecture
  and data management 62–63
  definition of xxi
  frameworks 63
  enterprise awareness 99
enterprise resource planning (ERP)
  software xxi, 8, 55, 132
  entities
    data definition of 86
    definition of xxi, 21–22
  entity integrity 108
  entity subtypes
    and structured types 150
    as alternative to exclusive relationship 71–72
    generally 67–69
    naming of 184
    notations supporting 163, 164
    entity supertypes 67
enterprise systems (ES) 163
  EPISTLE (European Process Industry STEP Technical Liaison Executive) 178–181
  exclusive relationships
    as alternative to subtypes 71–72
    notation for 69–70
eXtensible Stylesheet Language (XSL) 211
extent 148
external level schemas 16
  extraction, translation and loading
  (ETL) processes 139
  extranet 156
file based data systems 11
filegroups 52
Finkelstein, C. 162
first-normal form (1NF) xxi, 38–40
foreign keys
  constraints 108
  declaring 20, 51, 151
  definition of xxi
  in relational model of data 40
  fragmentation 134–137
  Full-Text package 155
  full replication 138
  function/procedure privileges 103
galaxy schema 145
generalisation hierarchies 67, 164
Genero, M.F. 98
global logical schemas 133, 137
granularity of data 144
Gregory, W.W. 98–99
Hay, D. 159
hierarchical model of data
  limitations of 205–211
  overview 169–171
horizonal fragmentation 135
HTTP (HyperText Transfer Protocol) 156
hybrid fragmentation 135
hypercubes 140
ICAM Definition Languages (IDEF) 163
  identifying relationships 160, 164
  inconsistency 5–6, 7
  incremental backups 110
  independent entities 164–165
  indexes 15, 52–53, 115
  information
    and data 2, 3–4
    as business resource 1–2, 8–9
    definition of xxi, 1, 2
    importance of high quality 4–5
    sharing xxix, 6, 9–10
  information management
    definition of xxi
    relationship with data management 60
Integration Management System (IMS) 169
information requirement analysis 17–19
inheritance
  and entity subtypes 68–69
  in object oriented programming 146
  in object relational databases 152
Integrated Database Management System (IDMS) 173
internal level schemas 15
Internet, the 155
intranet 156
JavaScript 157, 205
network model of data 172–174
neural networks 143
non-identifying relationships 164
normal form
Boyce-Codd-normal form xix, 47–48
definition of xxi
first-normal form xxii, 38–40
second-normal form xxiii, 40–44
third-normal form xxiv, 35, 44–47
normalisation xxii, 35–47, 138
NULL 19
object-oriented databases 15, 147–149
Object Data Management Group
(ODMG) specification
attributes 148
extent 148
keys 148
methods 149
Object Definition Language (ODL) 147
Object Query Language (OQL) 147
relationships 148–149
Object Definition Language (ODL) 147
object identifier 145, 147, 150–152
object orientation
aggregation 146
behaviour 145
class 145
definition of xxi, 145
encapsulation 146
inheritance 146
methods 145
polymorphism 146–147
state 145
Object Query Language (OQL) 148
object relational databases
collection types 148
structured types 149–152
objects 145
occurrence data, quality of 97, 98
off-the-shelf packages 131–133
Olson, J.E. 95
one-to-many relationships 25
one-to-one relationships 28
online analytical processing (OLAP)
xxii, 142–143, 145
online transactional processing (OLTP)
xxii, 142
organisations, and data management 124–130
ownership
of data definition xx, 85
of data values xx, 85
packages
for multimedia data 155
off-the-shelf 131–133
partial replication 138
partitioning 135–137
performance
and design 113
causes of poor performance 107, 114
improving 16, 52–53, 114–115
PHP 157, 205
physical data independence 12, 16, 53
physical file storage 52
physical redundancy 110–111
Piattini, M.G. 98
policies
data management 58
data recovery 109
data security 102
polymorphism 146–147
primary keys
constraints 108
declaring 20, 51
definition of xxi
in relational model of data 35, 38
multiple-part 38–40, 41
primary names 86, 184
prime terms 89, 193
privileges 102–104
qualifications, data management xxix–xxx
Qualified Dublin Core 92
query languages
Object Query Language (OQL) 147
XQuery 211
recursive relationships 23
redundancy of data 110–111
referential integrity 108
Reingruber, M. 98–99
relational data analysis xxiii, 33–47
relational database management systems (RDBMS) xxii–xxiii
relational databases
extracting data from XML 217
representing data as XML 213–217
relational model of data
attributes 34
definition of xxi
description of 33, 169
tuples xxiv, 34
relations
composition of 33
definition of xxi
normalisation of 35–47
rules for 34
relationships
between relations 40
definition of xxi
exclusivity 69–72
identifying 160, 164
in conceptual data models 23–25
many-to-many 29
resolving 29
non-identifying 164
one-to-many 25
one-to-one 28
recursive 23
repeating groups 38
replication 111, 137
repositories
as centralised information source 118, 120–121
definition of xxi, 195
features of 118–120
procurement of 119–120
purpose of 116–118
repository administration
definition of xxi, 61–62
position in organisation 125–130
responsibilities of 116–121
Index

skills 125
requirements, analysis of 16–19
Resource Description Framework 92
restricted terms, in naming 191–193
roll-up (aggregation) operation 142
rows 14

schemas
definition of xxiii, 12
for data warehouses 144–145
for distributed databases 133, 135, 137–138
scripting languages
JavaScript 157
PHP 157, 205
scripts (SQL) 19
second-normal form (2NF) xxiii, 40–44
sequential access 11
SGML (Standard Generalised Markup Language) 155, 205
shredding 217
significance statements 85
skills
data administration 121–125
database administration 123–125
repository administration 125
slice operation 142
snowflake schema 145
Spatial package 155
SQL
and multimedia data 152–155
and object relational databases 149–152
and OLAP operations 145
conceptual view of data 14
data control language 102
data definition language 102
data manipulation language 102
foreign keys 20, 51, 151
generally xxiii, 13
primary keys 20, 51
privileges 102–104
scripts 19
SQL/XML
features of 156, 213
mapping from relational database 213–217
shredding 217
XML DATATYPE 156, 213, 215
XML publishing functions 214–217
XMLSERIALIZE function 217
XMLTABLE function 217
standby databases 110
star schema 144–145
state, of object 145
statistical techniques
in data mining 143
Still Image package 155
structured data xxiii, 4, 152
structured types 149, 151–152
for multimedia data 169
supertypes, entity 67
support threshold 201
surrogate keys 48, 177
synchronous replication 138
synonyms 85
syntactic completeness 99
syntactic correctness 99
't system high ' 106
table privileges 103
tables
allocation of 52
creation of 19
definition of xxiii, 14
naming of 51, 189
tablesspaces 52
tags
HTML 155, 205
XML 208
thesaurus 90
three-normal form (3NF) xxiv, 44–47
three-level schema architecture 15–16
TIFF format, metadata in 93
timeout interval 115
TOGAF 63
top-down distribution 133, 135–137
Total Information Quality Management (TIQM) 100
Total Quality data Management (TQdM) 100
trade-off triangle 82
transaction logs 109–111, 115
transactions
ACID properties 109
failures of 109
stored in data warehouse 139
transitive dependency 44
tuples xxiv, 34
UML (Unified Modelling Language)
as alternative to data modelling 18
overview xxiv, 18, 166–167
unstructured data xxiv, 4, 152
update-anywhere updating 138
updating, in distributed databases 138
URI (Uniform Resource Indicator) 155
validation criteria 85
valid operations 85
valid values 85
vertical fragmentation 135
views, for access control 103, 104–105
web browsers 155, 156, 205
web servers 155, 157
web technology, and databases 155–157
West, M. 179
World Wide Web, the 155
XLink 211
XMLAGG function 215
XML datatype
from XML publishing functions 214, 215
uses 156, 213
XMLELEMENT function 215–216
XML (eXtensible Markup Language)
as data source for relational database 156–157, 213–217
common element definitions 210
components of 210
criteria for valid document 209–211
definition of xxiv
DTD (Document Type Definition) 210
extracting from relational databases 156–157, 217
native XML databases (NXDs) 156
tags 208, 209
XMLFOREST function 215
XML publishing functions 214–217
XML Schema Definition (XSD) 211
XMLSERIALISE function 217
XMLTABLE function 217
XPath 211
XPointer 211
XQuery 156, 211, 217
XSD (XML Schema Definition) 211
XSL (eXtensible Stylesheet Language) 211
XSLT(XSL Transformations) 211
Zachman Framework 63