Preface

We live in an environment where data (information in binary digital form) surrounds us and is essential for most activities in which we participate. Librarians and archivists act as data creators, data users and reusers, and/or data curators in increasingly digitally oriented environments. Despite this, professional practice has not caught up with digital practice in many respects. Caring for data, ensuring its usability and reuse in the future, and ensuring its accessibility and understandability over time require new strategies, practices, and tools. Traditional library and archival practices developed in a predigital and largely paper-oriented environment do not automatically transfer to the current digitally oriented environments. Although the past decade has seen the rapid development of new strategies, practices, and tools, these are not yet sufficiently mature. Consider these facts:

- Immense quantities of information in binary digital form are being generated in all walks of life.
- The quantities are increasing at a rapid rate.
- The scientific, scholarly, and research communities increasingly rely on networked computing, as trends such as the move from *in vitro* to *in silico* science and the development of large digital libraries in the humanities become dominant.
- Computer technology (hardware, software, and communications networks) quickly becomes obsolete.

All of these place data at risk from factors such as technology obsolescence, digital object fragility, a lack of understanding about what constitutes good practice, insufficient resources, and inappropriate organizational infrastructure.

Although the body of practice known as *digital preservation* is developing to address the factors that place data at risk, it is starting to be commonly accepted that its outcomes provide only part of the answer. For example, it is relatively straightforward to maintain a bit stream over time: there are more than 40 years of practice to call on in this respect. However, there is no guarantee that the data represented in this bit stream have the characteristics that allow them to be used and understood in the future, and to remain unchanged. How can these characteristics be retained in the data that is maintained for use in the future?

To answer this question effectively, more than simply a focus on maintaining the data (i.e., digital preservation) is required. What must also be considered is what comes before preservation and what comes after—that is, how the data are created and how they are used before they get to an archive or library and how they will be used, and by whom, in the future. This requires a focus on data that differs from that applied to physical artifacts such as books, manuscripts, and photographic prints in a predigital environment.

Digital curation is a developing set of techniques that address these issues, emphasizing the maintenance of data and adding value to these data for current and future use. Because it is still developing, digital curation is not yet described in detail in the literature. *Digital Curation: A How-To-Do-It Manual* therefore makes a significant contribution by describing in detail, in one place, the basics and current practices of digital curation.

Various models of the lifecycle of data are available. These typically begin with the creation of data and move through its various stages, ending with data use. The *Curation Lifecycle Model* developed by the Digital Curation Centre (DCC) (DCC, 2008; Higgins, 2008) is one of these. It was developed by the DCC to describe the processes involved in digital curation. The DCC Curation Lifecycle Model encompasses data from their conceptualization and creation through all aspects of their selection, archiving, maintenance, and use, to their reuse in the future. The DCC Curation Lifecycle Model provides an action-oriented structure for this book.

Digital Curation: A How-To-Do-It Manual is intended for anyone who creates data, anyone who uses and reuses data, and anyone who curates data. In essence, this means everyone who uses computers. More specifically, this book is intended to be read by librarians and archivists and by students of these professions. It should also have wider appeal, for example, to scientists and scholars who plan research and collect and use data. Whoever its readers, it will assist them to incorporate curation procedures, where relevant, into their own practice, figure out where to start when developing and implementing digital curation processes, and explore digital curation issues by providing a context for digital curation.

Digital Curation: A How-To-Do-It Manual is designed to be read in several ways. Its chapters can be read consecutively as an overview of digital curation, or they can be dipped into for general background and for advice on specific actions. The book's accompanying website (www.neal-schuman.com/curation) provides checklists that can be used separately as guidance and reminders about the tasks that comprise digital curation actions and templates that can be downloaded and used as the basis for developing digital curation plans and procedures for specific libraries, archives, and other organizations, as well as providing guidance for informing individual practice.

This book is based on the author's extensive international experience as a researcher, author, and presenter in the field of digital preservation and digital curation. It draws in particular on his experience with digital curation in Australia (which is widely acknowledged as representing international best practice) and in the European Union context (including a period based at the Humanities Advanced Technology and Information Institute [HATII] at the University of Glasgow), and his current work at Simmons College in Boston. The book's content was developed through observing digital curation practice, attendance at relevant international conferences, developing material for the DCC, and investigating the real-life experiences of digital curators, especially in the United States. It is also informed by a series of in-depth interviews with digital preservation professionals, which the author carried out as part of the preparation of his book *Preserving Digital Materials* (Harvey, 2005).

Examples of digital curation practice from the United Kingdom and Europe are well represented, as are examples from the United States and from other countries. It may on the surface seem surprising to the American reader that there are not more examples from the United States. There are three main reasons for this. The first is that digital curation is highly international and collaborative to an extent that is perhaps unprecedented in library science and archival practice (as noted in more detail in Chapter 8). This means that developments and practice in the field in one country are keenly observed and adopted, with modifications to suit local requirements, in other countries. The second is that, as Jordan and his colleagues note, "U.S. funding dedicated to digital preservation has traditionally lagged behind that available in the European and British contexts in particular" (Jordan et al., 2008). This means that the large majority of documented examples of digital curation developments and practice have to date come from outside the United States. This is likely to change in the next two to three years as funding such as the National Science Foundation's DataNet program comes on stream. The third is that the more centralized U.K. and European environments have required freely available documentation of any digital curation activities funded by public money. This point applies in particular to the material available on the DCC's website, which represents the only public documentation of a prolonged effort to identify and describe digital curation and to investigate practice in the field. This book therefore makes heavy use, with the permission of the DCC, of the materials accessible through the Centre's website.

Organization

Digital Curation: A How-To-Do-It Manual is organized in three parts. "Part I. Digital Curation: Scope and Incentives" provides a broad context for digital curation by introducing the main concepts and providing an overview. Chapter 1 indicates the reasons why digital curation is necessary, identifies what digital curation encompasses, suggests why one should be interested in digital curation, notes the main incentives for digital curation, and examines who does digital curation and what tasks they carry out. Chapter 2 notes the changing landscape in which librarians, archivists, researchers, and scholars work, its requirements for different ways of working and new kinds of infrastructure, and the different skill sets for data curation. Chapter 3 examines an important conceptual model for digital curation, the DCC Curation Lifecycle Model, on which this book is based, and a key standard, the OAIS Reference Model. Chapter 4 investigates in more detail what is meant by the term *data* and other related terms. This is important to think about because it allows us to address better an important question—What exactly is it that we want to curate?

"Part II. Key Requirements for Digital Curation" examines the DCC Curation Lifecycle's Full Lifecycle Actions—the essential basic requirements for all aspects of digital curation, which apply to all of the Sequential Actions noted in Part III of this book. Chapter 5 covers *Curate and Preserve*, one of four Full Lifecycle Actions, noting how digital preservation and digital curation differ, examining the aims of digital curation, and describing how these aims are achieved. Chapter 6 examines another Full Lifecycle Action, *Description and Representation Information*, the metadata and other information required for effective data curation. Chapter 7 notes the essential nature of planning and policy in data curation by describing a third Full Lifecycle Action, *Preservation Planning*. Chapter 8 completes the examination of Full Lifecycle Actions by describing *Community Watch and Participation* and noting the high value placed in digital curation on sharing knowledge and on collaboration.

"Part III. The Digital Curation Lifecycle in Action" is based on the DCC Curation Lifecycle's Sequential Actions and also notes its Occasional Actions. Chapter 9 notes the Sequential Action "Conceptualise," stressing the need to think about curation at the very first stages of planning research or creating digital objects. Chapter 10 examines the second Sequential Action, *Create or Receive*, noting the requirements for curation-ready digital objects. Chapter 11 describes *Appraise and Select*, the third Sequential Action, noting the importance of selection of the digital objects to be curated. This chapter also notes the Occasional Actions *Reappraise and Dispose*. The fourth Sequential Action, *Ingest*—the actions required when digital objects are taken into an archiving system—is the topic of Chapter 12. Chapter 13 discusses the preservation strategies and actions associated with *Preservation Action*, the fifth Sequential Action. Also included in this chapter is the Occasional Action *Migrate*. Chapter 14 focuses on the sixth Sequential Action, *Store*, which is concerned with what is required to provide acceptable data storage in the archiving system. Chapter 15 notes the seventh Sequential Action, *Access, Use, and Reuse*, examining the requirements for successful sharing and reuse of data in the future. It also notes the eighth Sequential Action, *Transform*, thus completing the data Curation Lifecycle Model.

A decade ago, little was known about how to assess and preserve the immense body of digitized material that can double in size in a matter of a few short years. Today, we have a body of international experience and expertise to draw on. *Digital Curation: A How-To-Do-It Manual* and its companion website (www.neal-schuman.com/curation) are designed as a comprehensive resource for best practices in this area. Preserving knowledge is a sacred trust; this resource will enable practitioners in all areas of human experience to better succeed at this crucial task.

References

Digital Curation Centre. 2008. "The DCC Curation Lifecycle Model." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed April 26, 2010).

Harvey, Ross. 2005. *Preserving Digital Materials.* Munich: K. G. Saur.

Higgins, S. 2008. "The DCC Curation Lifecycle Model." *International Journal of Digital Curation* 3, no. 1: 134–140. Available: www.ijdc.net/index.php/ijdc/article/viewFile/69/48 (accessed April 26, 2010).

Jordan, Christopher, Ardys Kozbial, David Minor, and Robert H. McDonald. 2008. "Encouraging Cyberinfrastructure Collaboration for Digital Preservation." Paper presented at iPres 2008, British Library, London, September 30, 2008. Available: www.bl.uk/ipres2008/presentations_day2/39_Jordan.pdf (accessed April 26, 2010).

Digital Curation: Scope and Incentives

I

The four chapters in "Part I. Digital Curation: Scope and Incentives" provide a broad context for digital curation by introducing the main concepts and giving an overview of the field.

Chapter 1 indicates the reasons why digital curation is necessary, identifies what digital curation encompasses, suggests why you should be interested in digital curation, notes the main incentives for digital curation, and examines who does digital curation and what tasks they carry out.

Chapter 2 notes the changing landscape in which librarians, archivists, researchers, and scholars work; its requirements for different ways of working and new kinds of infrastructure; and the different skill sets for data curation.

Chapter 3 describes the application of lifecycle models to digital curation and looks in more detail at a key conceptual model and a key standard for digital curation. The first, the Digital Curation Centre (DCC) Curation Lifecycle Model, outlines the actions that comprise digital curation and presents these actions in graphic form. This Lifecycle Model is used as the structural basis of Parts II and III of this book. The second lifecycle model, the Open Archive Information System (OAIS) Reference Model, is widely used as the basis for the design and implementation of digital archival systems.

Chapter 4 notes in more detail the meaning of the term *data* and of other related terms. Investigating the meaning of the term *data* is particularly important if a key question is to be answered satisfactorily: What exactly is it that we want to curate?

Introduction

1

IN THIS CHAPTER:

✔ Why There Is a Need for Digital Curation

✔ What Digital Curation Is

✔ Why We Should Be Interested in Digital Curation

✔ Incentives for Digital Curation

✔ Digital Curators

✔ Summary: Main Characteristics of Digital Curation

✔ References

Chapter 1 sets the scene for digital curation and argues that it is central to professional practice in all digital environments. It begins by indicating why digital curation is necessary, then identifies what it encompasses, briefly defines terms such as *data*, *digital object*, and *database* in this context, suggests why an interest in digital curation is important, notes the main incentives for digital curation, and examines the tasks that comprise digital curation and who carries them out.

Some definitions set the scene. First is a short working definition of digital curation. Digital curation is defined briefly by the Digital Curation Centre (DCC) as

> maintaining and adding value to a trusted body of digital research data for current and future use; it encompasses the active management throughout the research lifecycle. (Digital Curation Centre, accessed 2010)

Definitions of the terms *data*, *digital object*, and *database* used in this book come from the DCC Curation Lifecycle Model, the model upon which the structure of this book is based (Digital Curation Centre, 2008). *Data* is "any information in binary digital form." This definition is intentionally very broad and extends beyond the narrow connection of the word with the outputs of scientific research. It includes *digital objects* and *databases*. *Digital objects* can be simple or complex. "*Simple digital objects* are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata. *Complex digital objects* are discrete digital objects, made by combining a number of other digital objects, such as Web sites." *Databases* are "structured collections of records or data stored in a computer system." These definitions of the terms *data*, *digital object*, and *database* and their implications are expanded on in Chapter 4.

Why There Is a Need for Digital Curation

The increasingly digital world that we all inhabit is changing the ways we work and play. It is a truism that this results in the generation of massive quantities of data in all areas of our lives. Furthermore, these quantities are increasing at significant rates. (This point is easy to illustrate. Consider the amount of personal data—word-processed documents, digital photographs, video files, and so on—you thought you needed to store ten, even only five, years ago, and compare that with the quantity you now think you need to store.) Data, whether personal or of any other kind, has certain characteristics that require it to be actively managed. It is at risk from many factors, including:

- technology obsolescence—computers and software are updated frequently, often resulting in inability to access data;
- technology fragility—digital objects can become inaccessible if only a small part of them is changed or corrupted;
- lack of understanding about what constitutes good practice—digital curation is a new and still-developing field of practice, and much about what is needed to make it work is still unknown;
- inadequate resources—libraries, archives, and museums are usually not resourced to carry out all they want to do; digital curation is not always given a high priority, and understanding what skill sets are required to make digital curation work is not fully known; and
- uncertainties about the best organizational infrastructures to achieve effective digital curation.

Digital curation is also necessary for many other reasons. Many of the current developments in the field—its practices, tools, storage facilities, and theoretical bases—are coming from the scientific, scholarly, and research communities. These communities have been rapidly accommodating new ways of working that rely increasingly on networked computing to link researchers and scholars around the world and to generate and share large—in some cases extremely large—data sets. Historians, for example, "ignore the future of digital data at their own peril" if they do not "ensure the future of their own scholarship" which involves new prospects such as "linking directly from footnotes to electronic texts" (Rosenzweig, 2003: paragraph 64). A researcher in the future will work differently:

> Not only will there be text, with hyperlinks to related literature or citations within the article, there will be links to the data reported within the article, through graphs, tables, illustrations, that will link to related datasets. (ARL Workshop on New Collaborative Relationships, 2006: 141)

This will only be possible if stable digital curation is achieved.

These trends are often described in the context of science as the move from *in vitro* to *in silico* science—broadly speaking, from laboratory-based science to science based on data and performed using computers. These new contexts are collectively termed in the United States as *cyberscholarship* and in other countries as *e-science* or *e-scholarship*. Chapter 2 describes these trends in more detail.

Cyberscholarship generates large quantities of data. This data is often unique and cannot be reproduced without major cost, if at all. An example is environmental data. The data may be generated in an extremely expensive experiment, and the cost alone means that the experiment cannot be reproduced: an example, perhaps extreme, is the massive amounts of data generated from runs of the world's largest and highest energy particle accelerator, the Large Hadron Collider.

Cyberscholarship also requires that data be available for use and for reuse in the future. There are many reasons. Large data sets can be the basis of analysis by scholars around the world, so they must be available for access. Good research and scholarship are based on data that can be verified and built on to lead to new knowledge. Data may be records that have legal requirements: for example, financial records of business transactions may be required to be kept for periods of time specified in legislation. Some funding agencies require that data created during the course of activities they fund be made available for public use and reuse. The long-acknowledged roles of libraries and archives in preserving social memory should also be noted as a significant reason for ensuring data are available for use and reuse in the future, as social memory is increasingly held in digital form.

One articulation of cyberscholarship is the U.K. Research Information Network's *Stewardship of Digital Research Data: A Framework of Principles and Guidelines* (Research Information Network, 2008: 3). The Framework's five principles are based on sharing and reusing research data. The principles indicate the need for international standards to be developed and applied to the creation and collection of data, the importance of making this data able to be located and easy to use, and the need to protect rights of data creators and owners—all with an emphasis on efficiency and cost-effectiveness. The last of these principles is: "Digital research data of long term value arising from current and future research should be preserved and remain accessible for current and future generations."

For all of these reasons, actively managing data over their lifecycle is essential. Digital curation is a set of techniques that address the issues of data protection and risk management to ensure that the data are available and usable now and in the future.

What Digital Curation Is

The brief working definition of digital curation noted at the beginning of this chapter comes from the DCC, the principal organization in the United Kingdom for developing and promoting digital curation concepts and practices. Another definition, this one from the United States, is provided by the Digital Curation Curriculum (DigCCurr) project based at the University of North Carolina at Chapel Hill in a description of its interests. It expands on the DCC's brief definition:

> Our cultural heritage, modern scientific knowledge, and everyday commerce and government depend upon the preservation of

> reliable and authentic electronic records and digital objects. While digital data holds the promise of ubiquitous access, the inherent fragility and evanescence of media and files, the rapid obsolescence of software and hardware, the need for well-constructed file systems and metadata, and the intricacies of intellectual property rights place all of these materials at risk and offer little hope of longevity for information that is not intentionally preserved. A decade of work in digital preservation and access has resulted in an emerging and complex life-cycle constellation of strategies, technological approaches, and activities now termed "digital curation." (DigCCurr, accessed 2010)

Being aware of where these definitions originated helps us to better understand the concerns of digital curation and its current emphases. There were two drivers to the establishment of the DCC in the United Kingdom: e-science, the "data deluge" and continuing access to the data sets generated; and digital preservation, particularly the realization that digital preservation activities were by themselves insufficient to address many of the issues associated with maintaining data over time. In the United Kingdom this resulted in the development by the Joint Information Systems Committee (JISC) of a "Continuing Access and Digital Preservation Strategy" (Joint Information Systems Committee, 2002) and Lord and Macdonald's (2003) report about data curation for e-science, one outcome of which was the release of funding in 2003 to establish a digital curation center. The DCC was established in 2004. Because of its basis in e-science, much of the data curation literature and activities in the United Kingdom were initially focused heavily on scientific data, although in recent years this scope has broadened. In general, the same can be said for the United States, where initial interest in the need for data curation came from the National Science Foundation. But the scope has recently been broadened considerably through the interest of groups such as the Research Libraries Group (now merged with OCLC), the Association of Research Libraries, and the National Endowment for the Humanities to include humanities and social science data. In both countries significant interest has also been expressed in the curation of personal data.

It is important to reinforce the last point: that significant interest has been shown in the curation of personal data. While it is true that most of the recent understandings and practices of digital curation have been developed for and by the scientific communities, much of it is highly applicable, often without modification, to all information in digital form, whether personal data or data preserved by libraries and archives. The reader is urged to keep this in mind when reading and applying the points noted to his or her own context or area of interest.

Just which of the "emerging and complex life-cycle constellation of strategies, technological approaches, and activities" (DigCCurr, accessed 2010) make up digital curation? This is understood differently by different groups. Some of the activities that make up digital curation are reported by Brophy and Frey (2006: 38):

- Maintaining the links between digital information and associated annotations or published materials, including citations

- Ensuring the long-term accessibility and reusability of digital information
- Performing archiving activities on digital information such as selection, appraisal, and retention
- Ensuring the authenticity, integrity, and provenance of digital information are maintained over time
- Performing preservation activities on digital information such as migration or emulation
- Maintaining hardware components to enable digital information to be accessed and understood over time
- Managing digital information from its point of creation
- Managing risks to digital information
- Ensuring the destruction of digital information

These are all aspects of digital curation, but this list does not present the whole picture. So we are still left with the question: what is digital curation? We can state what digital curation is *not*:

1. It is not *digital archiving*—one definition of digital archiving is "the process of backup and ongoing maintenance as opposed to strategies for long-term digital preservation" (Digital Preservation Coalition, 2008: 24).
2. It is not *digital preservation*—defined as "all of the actions required to maintain access to digital materials beyond the limits of media failure or technological change" (Digital Preservation Coalition, 2008: 24) and as "policies, strategies and actions that ensure access to digital content over time" (ALCTS Preservation and Reformatting Section, 2007).

Although digital archiving and digital preservation are important aspects of digital curation, they are not the whole story. Lavoie and Dempsey (2004) describe the position:

> Our understanding of the totality of the challenges associated with maintaining digital materials over the long-term is coming more sharply into focus. New questions are emerging, having less to do with digital preservation as a technical issue *per se*, and more to do with how preserving digital materials fits into the broader theme of *digital stewardship*. These questions surface from the view that digital preservation is not an isolated process, but instead, one component of a broad aggregation of interconnected services, policies, and stakeholders which together constitute a digital information environment.

Digital curation is a more inclusive concept than either digital *archiving* or digital *preservation*. It addresses the whole range of processes applied to digital objects *over their lifecycle*. Digital curation begins before digital objects are created by setting standards for planning data collection that results in "curation-ready" digital objects that are in the best possible condition to ensure they can be maintained and used in the future. Digital curation emphasizes *adding value* to data sets and digital objects, through things such as additional metadata or annotations, so that they can be reused. Digital curation involves a *wide range of stakeholders* cutting across disciplinary boundaries: as well as cultural heritage organizations such as libraries, archives, and museums, it also involves funding agencies, government bodies, national data centers, institutional repositories, and learned societies. (In fact, digital curation is the concern of all who create and use data.) Digital curation is also concerned with *risk management*: it "is about converting uncertainties into measurable and manageable risks" (DRAMBORA, 2007). It is also about *good data management* practices.

Digital curation is concerned with and applicable to a wide range of digital objects. It is as equally applicable to complex digital objects that are *linked* to other resources in a range of formats, large science data sets, or data sets that are changing every second, as it is to relatively simple digital objects such as the static documents usually handled by libraries and archives. However, most data archiving and digital preservation practices were developed for static documents; they do not transfer successfully to more complex data. Although professional attention has been paid to digital collections in libraries and archives for many years (digital library activities are a case in point), it has typically focused on only part of the lifecycle, usually digitizing and providing access to the digitized information. Such actions cannot be considered as sufficient for digital curation, which is concerned with the whole lifecycle and emphasizes maintaining digital information over time and ensuring its availability and usability in the future. In this new era of data-driven scholarship and research, new strategies and processes are needed to handle the wide range of data created and maintained by many different kinds of user communities.

Taking all of this into account, an expanded definition of digital curation might read: Digital curation is concerned with actively managing data for as long as it continues to be of scholarly, scientific, research, administrative, and/or personal interest, with the aims of supporting reproducibility, reuse of, and adding value to that data, managing it from its point of creation until it is determined not to be useful, and ensuring its long-term accessibility, preservation, authenticity, and integrity.

Why We Should Be Interested in Digital Curation

That digital curation is necessary and a matter of urgency is generally understood by anyone who uses computers. Seamus Ross (2007: 2), a prominent researcher in several areas of digital curation, describes the reasons why digital objects and data become unusable:

> They are bound to varying degrees to the specific application packages (or hardware) that were used to create or manage them. They are prone to corruption. They are easily misidentified. They are generally poorly described or annotated.... Where they do have sufficient ancillary data, these data are frequently time constrained.

Figure 1.1 lists the threats to digital continuity—that is, to the continuing accessibility and usability of data. The figure clearly indicates the most significant reasons why digital curation is an urgent imperative.

Obsolescence is probably the most commonly recognized of these threats. Our abilities to maintain digital objects and to use them over time are challenged by the wide range of formats, of both software and hardware, and by their rapid rates of change. Examples abound, among them the fact that personal computers are no longer supplied with a drive to read and write three-and-a-half inch diskettes, which were only a few years ago the standard data storage medium for personal use. Some of the wide range of storage media and computer formats are displayed in online exhibits. Two of these are:

1. *Timeline: Digital Preservation and Technology* and *Chamber of Horrors: Obsolete and Endangered Media*—accessed through the introduction to the Cornell University Library's online tutorial *Digital Preservation Management* (Cornell University Library, 2003–2007)

Figure 1.1. Threats to Digital Continuity

The carriers used to store... digital materials are usually unstable and deteriorate within a few years or decades at most
Use of digital materials depends on means of access that work in particular ways: often complex combinations of tools including hardware and software, which typically become obsolete within a few years and are replaced with new tools that work differently
Materials may be lost in the event of disasters such as fire, flood, equipment failure, or virus or direct attack that disables stored data and operating systems
Access barriers such as password protection, encryption, security devices, or hardcoded access paths may prevent ongoing access beyond the very limited circumstances for which they were designed
The value of the material may not be recognised before it is lost or changed
No one may take responsibility for the material even though its value is recognised
Those taking responsibility may not have adequate knowledge or facilities
There may be insufficient resources available to sustain preservation action over the required period
It may not be possible to negotiate legal permissions needed for preservation
There may not be the time or skills available to respond quickly enough to a sudden and large change in technology
The digital materials may be well protected but so poorly identified and described that potential users cannot find them
So much contextual information may be lost that the materials themselves are unintelligible or not trusted even when they can be accessed
Critical aspects of functionality, such as formatting of documents or the rules by which databases operate, may not be recognised and may be discarded or damaged in preservation processing

Source: Guidelines for the Preservation of Digital Heritage, March 2003. © UNESCO 2003. Used by permission of UNESCO.

2. The Computer History Museum's virtual exhibit *Timeline of Computer History* (Computer History Museum, 2006)

The increasing quantities of data produced in digital form and their increasingly dynamic nature (exemplified by large online databases that are continually being added to by contributors around the world) pose another major threat to digital continuity, challenging our ability to capture, store, and access these data. The increasing quantities also demand that decisions are made about which data to curate, as not all data are created equal. This raises challenging questions such as: How do we decide what is likely to be useful in the future? Useful to whom? How long should we plan to keep them? Do we want them to be usable (functional), and to what extent, in the future?

Responses to threats to digital continuity that are based on traditional preservation approaches do not work. Simply capturing data on stable storage media and copying them onto new storage media when obsolescence threatens are in themselves not sufficient to ensure digital continuity. Digital data must be managed from the point that they are created (or, ideally, before they are created) if their survival is to be ensured. Active management of data over the whole of their life is necessary, requiring "constant maintenance and elaborate 'life-support' systems" (Hedstrom, 2002). Social and institutional issues must also be addressed: where, for example, does the continuing funding come from to maintain data in a research environment that is project oriented? This book identifies responses to these challenges.

An analysis of the curation of research data in Canada in 2008 provides a snapshot of the current situation and indicates clearly that there is cause for alarm (Research Data Strategy Working Group, 2008). Using a four-part data lifecycle framework (data production, data dissemination, long-term data management, data discovery and repurposing) and ten indicators (policies, funding, roles and responsibilities, standards, data repositories, skills and training, accessibility, and preservation), this analysis assessed Canada's current state against an "ideal state" based on existing international best practice. The conclusion was that major barriers exist to accessing and preserving research data in Canada, with significant implications for the future of Canadian research and innovation. For example, large amounts of data are currently being lost because Canada does not have enough trusted data repositories. The following main issues in the curation of research data in Canada were identified:

- Data Production
  - Priority is on immediate use, rather than potential for long-term exploitation.
  - Limited funding mechanisms to prepare data appropriately for later use.
  - Few research institutions require data management plans.
  - No national organization that can advise and assist with application of data standards.
- Data Dissemination
  - Lack of policies governing the standards applied to ensure data dissemination.

  - ○ Researchers unwilling to share data, because of lack of time and expertise required.
  - ○ Some policies require certain types of data be destroyed after a research project is over.
- Long-Term Management of Data
  - ○ Lack of coverage and capacity of data repositories.
  - ○ Preservation activities in repositories are not comprehensive.
  - ○ Limited funding for data repositories in Canada.
  - ○ Few incentives for researchers to deposit data into archives.
- Discovery and Repurposing
  - ○ Most data rests on the hard drives of researchers and is inaccessible by others.
  - ○ Per [i.e., pay] per view and licensed access mechanisms are common where data are available.
  - ○ Many researchers are reluctant to enable access to their data because they feel it is their intellectual property. (Research Data Strategy Working Group, 2008: 16)

Canada is by no means alone in facing significant barriers in curation of research data. The Canadian report notes similar issues in the United States, the United Kingdom, Australia, and elsewhere.

Another cause for alarm is expressed in a 2008 survey of the preparedness for digital preservation of local governments in the United Kingdom. Over 80 percent of the respondents already held digital records. Although nearly half had a digital preservation policy, had undertaken some planning, and gave high priority to preserving digital records, awareness of the issues was low. Barriers to digital preservation were identified as cultural ("organisation, political, awareness, external partnerships/relations and motivation"), resource ("time, costs, funding, storage"), and skills gap ("Training, competencies, IT") (Boyle, Eveleigh, and Needham, 2008). If digital curation practice in this sector is not addressed as a matter of some urgency, there will be crucial losses of data.

The situation is not, however, as uniformly bleak as some commentators would lead us to believe. The issues were initially described and promoted in alarmist terms, to the extent that the term "digital dark age" has entered the collective consciousness through a Wikipedia entry (Wikipedia, 2009; Harvey [2008] provides other examples of alarmist terms and their consequences). But, as Lavoie and Dempsey (2004) remind us, "accumulating experience in managing digital materials has tempered this view."

Incentives for Digital Curation

To date, much of the money for digital preservation and digital curation has been short-term project-based funding. This project-based funding model does not support good digital curation practice. Because of the finite time span of projects, employees focus on their next job application or on getting funding for the next project. In this context a high priority is not usually placed on getting the data in good shape for curation beyond the life of the project. For example, there is often a lack of metadata to describe the data so that they are understandable. Data curation tasks are "that extra burden, the one just beyond what is currently possible, in the queue behind meeting the conference deadline and writing the grant application" (Rusbridge, 2007: 4). In these contexts it is important to be clear about how data curation is of benefit so that continuing interest in and application of digital curation are encouraged and maintained.

In an environment of competing priorities and multiple demands on our time, why should we be interested in the curation of data? The answer is clear: curation has immediate and short-term benefits for all who create, use, and manage data, in four main ways:

1. **Improving access.** Digital curation procedures allow continuing access to data and improve the speed of access to reliable data and the range of data that can be accessed.
2. **Improving data quality.** Digital curation procedures assist in improving data quality, improving the trustworthiness of data, and ensuring that data are valid as a formal record (such as use as legal evidence).
3. **Encouraging data sharing and reuse.** Digital curation procedures encourage and assist data sharing and use by applying common standards and by allowing data to be fully exploited through time (thus maximizing investment) by providing information about the context and provenance of the data.
4. **Protecting data.** Digital curation procedures preserve data and protect them against loss and obsolescence.

Digital curation does all of this by providing tools and services to migrate digital objects plus their associated metadata into new formats that stay meaningful to users and by providing a management infrastructure for preserving them over time.

The benefits of participating in digital curation can be considered in three categories: direct benefits to data creators, "public good" obligations (such as the increasing interest in open access), and compliance reasons.

Direct Benefits to Data Creators

Good digital curation practices benefit data creators in many ways: improved quality of data, improved access to data, increased visibility of the research, and improved visibility and citation rates of the creator. Good digital curation practices also result in improved risk management, meaning that digital objects are more likely to remain usable over time. Examples of risks related to data, as noted earlier, include failure of storage media, hardware, or network services; obsolescence of media, hardware, and software; economic failure resulting in insufficient funding to maintain data over the long term; and organizational failure, where the parent organization no longer sees itself in the digital archiving business and wishes to dispose of its data. Risk management methodologies assist with developing lists of potential risks, assessing the likelihood of them occurring, and identifying their potential impact. These form the basis of policies and procedures to minimize the likelihood of risky events occurring and to manage risks.

"Public Good" Obligations

Some incentives for digital curation relate to public good. Pressure is increasingly being brought to bear to make data more broadly available for public scrutiny by community groups, for example, taxpayers' groups.

The Open Access movement is an example of the acknowledgment of "public good" obligations. The aim of open access is the free and unrestricted online availability of research results—a typical definition of it is "free, immediate, permanent online access to the full text of research articles for anyone, webwide" ("Open Access," accessed 2010). Participation in open access initiatives can assist data creators such as researchers and scholars to maximize their research impact. (A bibliography on the Open Citation Project [accessed 2010] website lists studies about the effects of open access on citation impact.) The return on public investment in research can also be maximized by reporting and citing that research more widely so that it forms the basis of further research; here, open access initiatives can assist. Research funding bodies are increasingly expecting open access to the research they fund. The Wellcome Trust, a major U.K.-based funder of medical research, has called for "Open and unrestricted access to the outputs of published research" (Wellcome Trust, accessed 2010).

Open access initiatives are gaining strength. A 2007 petition to the European Commission ("Petition for Guaranteed Public Access to Publicly-Funded Research Results," 2007) urges the adoption, as a matter of urgency, of a recommendation to guarantee public access to publicly funded research results shortly after publication. Open access journals are firmly established; for example, the Public Library of Science (PLoS, accessed 2010) is a library of open access journals and other scientific literature: "Everything we publish is freely available online for you to read, download, copy, distribute, and use (with attribution) any way you wish." The strength of the Open Access movement can be seen in the Directory of Open Access Journals (DOAJ, accessed 2010).

Compliance Reasons

Digital curation can also be compliance driven. Commonly encountered examples are compliance with the requirements of funding bodies and of publishers and the need to comply with specific legal requirements.

Research funding bodies now commonly require that grant applications include provision for digital curation. A data management plan, or a plan for the deposit of data into a publicly accessible data repository, is a common example. The National Institutes of Health (NIH) in the United States illustrates this point. "Data sharing is essential for expedited translation of research results into knowledge, products and procedures to improve human health," begins the NIH's data-sharing policy

(National Institutes of Health, 2007). The NIH criteria for peer reviewing of grant applications include an expectation that data will be shared. Their statement "Access to Research Data" (National Institutes of Health, 2003) defines research data and outlines the process of seeking access. The NIH provides a ***Data Sharing Workbook*** (National Institutes of Health, 2004a). Testimonials on the NIH website (National Institutes of Health, 2004b) indicate the benefits of data sharing, such as more rapid availability of data and higher take-up and reuse rates. In the United Kingdom, deposition of data in existing databases or repositories, which are sometimes prescribed, is mandated. For example, the U.K. Economic and Social Research Council (ESRC) specifies the Economic and Social Data Service (2003–2009) repository, and the U.K. National Environment Research Council (NERC) specifies the NORA repository (NERC Open Research Archive, 2009).

Compliance with legislation may necessitate good digital curation practice. Many countries have data protection acts and freedom of information acts. Discipline-specific compliance requirements may also determine practice. In the United Kingdom, the Freedom of Information Act (2000), the Data Protection Act (1998), and the Environmental Information Regulations (2004) mandate requirements for data that require careful curation. Natural environment research in the United Kingdom, for example, may have to comply with the Antarctic Treaty; data sets may contain "environmental information" that falls within the definition of the Environmental Information Regulations 1992; a contract or Memorandum of Understanding with another body may specify what can and cannot be done with the data. Details of these examples can be found in the *NERC Data Policy Handbook* (Natural Environment Research Council, 2002, Section 3.5).

In some disciplines publishers now insist that potential authors demonstrate aspects of digital curation. The publisher may require specific conditions to be met before publication of research results, such as registering clinical trials in a publicly accessible database as a precondition of publication—this is the case for major medical journals, such as the *British Medical Journal*, the *Journal of the American Medical Association*, the *New England Journal of Medicine*, and *The Lancet.*

Digital Curators

The creators, users, and curators of data all play roles in the digital curation process. The roles range from those of curators of large data sets in scientific, library, and archive contexts, right down to those played by individuals who create and use digital information for personal use and who wish to keep some of it over time.

Creators of data include scholars, researchers, and librarians and archivists who manage digitization programs. The best time to ensure that digital objects are usable is when they are created. For these objects to be usable and reusable, they must be of high quality, well structured, and adequately documented. Data creators, therefore, should ensure that the digital objects they create are structured and documented to ensure their longevity and reusability. Data reusers ensure that any annotations they produce are captured and documented to a level that ensures their annotations are understandable to other users of those data.

Curators of digital information—people who have a primary role of managing or "looking after" data—have job titles that include archivist, librarian, data librarian, and annotator, as well as data curator. Their roles vary according to the context in which they work. For example, in a bioscience context the data curator's tasks include ongoing data management, intensive data description, ensuring data quality, collaborative information infrastructure work, and metadata standards work.

The DCC's website provides case studies that describe what curation actually involves in practice (www.dcc.ac.uk/resources/case-studies). Among the full range of tasks and responsibilities encompassed by digital curation are these:

- Developing and implementing policies and services
- Analyzing digital content to determine what services can be provided from it
- Providing advice to data creators and users/reusers
- Ensuring submission of data to a repository
- Negotiating agreements
- Ensuring data quality
- Ensuring that data are structured in the best way to provide access, rendering, storage, and maintenance
- Enabling the use and reuse of data
- Enabling data discovery and retrieval
- Preservation planning and implementation (e.g., ensuring appropriate storage and backup routines, obsolescence monitoring)
- Ensuring that policies and services are in place to make sure that data is viable, able to be rendered, understandable, and authentic
- Promoting interoperability

Summary: Main Characteristics of Digital Curation

Digital curation is characterized by:

- the range of processes applied to digital objects *over their whole lifecycle*, from creation to ultimate disposal (e.g., it places strong emphasis on the importance of designing for curation at the point that digital objects are created);
- a concern with reproducibility of data as the basis of validation of scholarly output, accountability, and recordkeeping;

- adding value to digital objects so that they can be reused or repurposed (e.g., by adding metadata that assists in their discovery, management, and retrieval);
- involving a wide range of stakeholders cutting across disciplinary boundaries: these include heritage organizations (libraries, archives, museums, art galleries), e-science and e-research groups, researchers and scholars, and government bodies who fund e-science, higher education, and other activities;
- a strong interest in open source solutions; and
- strong links between research and practice.

Our understanding of digital curation is evolving. This becomes clear when we attempt to apply current digital curation practices to the e-science context. Much current digital curation practice has been developed in cultural heritage contexts, libraries and archives in particular, and is most effective for static data. This does not transfer readily to the new scholarship based on collaborative computing. This new scholarship is evolving very rapidly, lacks standards, and deals with very large data sets. There is a huge potential for reuse of data, but the infrastructure components to allow this reuse are currently very primitive or—more likely—do not yet exist. The next chapter examines the new ways of working, their requirements for digital curation, and the need to develop new kinds of skills.

References

ALCTS Preservation and Reformatting Section. 2007. "Definitions of Digital Preservation." Chicago: Association for Library Collections & Technical Services (June 24, 2007). Available: www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf (accessed April 26, 2010).

ARL Workshop on New Collaborative Relationships. 2006. "To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering: A Report to the National Science Foundation from the ARL Workshop on New Collaborative Relationships: The Role of Academic Libraries in the Digital Data Universe, September 26–27, 2006, Arlington, VA." Washington, DC: Association of Research Libraries. Available: www.arl.org/bm~doc/digdatarpt.pdf (accessed April 26, 2010).

Boyle, Frances, Alexandra Eveleigh, and Heather Needham. 2008. "Report on the Survey Regarding Digital Preservation in Local Authority Archive Services." York: Digital Preservation Coalition (November 3, 2008). Available: www.dpconline.org/docs/reports/digpressurvey08.pdf (accessed April 26, 2010).

Brophy, Peter, and Jeremy Frey. 2006. "Digital Curation Centre Externally-Moderated Reflective Self-Evaluation: Report." Edinburgh: Digital Curation Centre. Available: ie-repository.jisc.ac.uk/198/1/dcc_evaluation_report_final.pdf (accessed April 26, 2010). Used by permission of Peter Brophy.

Computer History Museum. 2006. *Timeline of Computer History.* Mountain View, CA: Computer History Museum. Available: www.computerhistory.org/timeline (accessed April 26, 2010).

Cornell University Library. 2003–2007. *Digital Preservation Management: Implementing Short-Term Strategies for Long-term Problems.* Ithaca, NY:

Cornell University Library. Available: www.icpsr.umich.edu/dpm/dpm-eng/eng_index.html (accessed April 26, 2010).

DigCCurr. Available: ils.unc.edu/digccurr/aboutI.html (accessed April 26, 2010).

Digital Curation Centre. "DCC Charter and Statement of Principles." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/about-us/dcc-charter (accessed April 26, 2010).

———. 2008. *The DCC Curation Lifecycle Model.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed April 26, 2010).

Digital Preservation Coalition. 2008. *Preservation Management of Digital Materials: The Handbook.* York: Digital Preservation Coalition. Available: www.dpconline.org/docs/advice/digital-preservation-handbook.html (accessed April 26, 2010).

DOAJ: Directory of Open Access Journals. Lund: DOAJ. Available: www.doaj .org (accessed April 26, 2010).

DRAMBORA: Digital Repository Audit Method Based on Risk Assessment. 2007. Edinburgh: DRAMBORA. Available: www.repositoryaudit.eu/img/drambora_flyer.pdf (accessed April 26, 2010).

Economic and Social Data Service. 2003—2009. Colchester: UK Data Archive. Available: www.esds.ac.uk (accessed April 26, 2010).

Harvey, Ross. 2008. "So Where's the Black Hole in Our Collective Memory? A Provocative Position Paper." Glasgow, Scotland: DigitalPreservationEurope. Available: www.digitalpreservationeurope.eu/publications/position/Ross_Harvey_black_hole_PPP.pdf (accessed April 26, 2010).

Hedstrom, Margaret. 2002. "Research Challenges in Digital Archiving and Long-term Preservation." Address to the Workshop on Research Challenges in Digital Archiving and Long-Term Preservation, Washington, DC, April 12–13, 2002. Available: www.sis.pitt.edu/~dlwkshop/paper_hedstrom.doc (accessed April 26, 2010).

Joint Information Systems Committee. 2002. "Continuing Access and Digital Preservation Strategy for JISC." Bristol: JISC (October 1, 2002). Available: www.jisc.ac.uk/publications/publications/pub_access_pres_strategy.aspx (accessed April 26, 2010).

Lavoie, Brian, and Lorcan Dempsey. 2004. "Thirteen Ways of Looking at... Digital Preservation." *D-Lib Magazine* 10, no.7/8 (July/August). Available: www.dlib.org/dlib/july04/lavoie/07lavoie.html (accessed April 26, 2010).

Lord, Philip, and Alison Macdonald. 2003. "e-Science Curation Report: Data Curation for e-Science in the UK: An Audit to Establish Requirements for Future Curation and Provision." Twickenham: Digital Archiving Consultancy. Available: www.jisc.ac.uk/uploaded_documents/e-ScienceReportFinal.pdf (accessed April 26, 2010).

National Institutes of Health. 2003. "Access to Research Data." In *NIH Grants Policy Statement.* Bethesda, MD: National Institutes of Health. Available: grants.nih.gov/grants/policy/nihgps_2003/NIHGPS_Part5.htm#_Access_to_Research (accessed April 26, 2010).

———. 2004a. *Data Sharing Workbook.* Bethesda, MD: National Institutes of Health. Available: grants1.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf (accessed April 26, 2010).

———. 2004b. "Testimonials." Available: grants1.nih.gov/grants/policy/data_sharing/testimonials.doc (accessed April 26, 2010).

———. 2007. *NIH Data Sharing Policy.* Bethesda, MD: National Institutes of Health. Available: grants1.nih.gov/grants/policy/data_sharing/index.htm (accessed April 26, 2010).

Natural Environment Research Council. 2002. *NERC Data Policy Handbook, Version 2.2.* Swindon: NERC. Available: badc.nerc.ac.uk/data/NERC_Handbookv2.2.pdf (accessed April 26, 2010).

NERC Open Research Archive (NORA). 2009. Swindon: Natural Environment Research Council. Available: www.nerc.ac.uk/about/access/repository.asp (accessed April 26, 2010).

"Open Access." eprints. Available: www.eprints.org/openaccess (accessed April 26, 2010).

Open Citation Project. "The Effect of Open Access and Downloads ('Hits') on Citation Impact: A Bibliography of Studies." Available: opcit.eprints.org/oacitation-biblio.html (accessed April 26, 2010).

"Petition for Guaranteed Public Access to Publicly-Funded Research Results." 2007. Available: www.ec-petition.eu/index.php?p=index (accessed April 26, 2010).

PLoS: Public Library of Science. San Francisco, CA: Public Library of Science. Available: www.plos.org (accessed April 26, 2010).

Research Data Strategy Working Group. 2008. *Stewardship of Research Data in Canada: A Gap Analysis.* Ottawa: National Research Council Canada. Available: data-donnees.gc.ca/docs/GapAnalysis.pdf (accessed April 26, 2010). Used by permission of the Research Data Strategy Working Group.

Research Information Network. 2008. *Stewardship of Digital Research Data: A Framework of Principles and Guidelines: Responsibilities of Research Institutions and Funders, Data Managers, Learned Societies and Publishers.* London: RIN. Available: www.rin.ac.uk/system/files/Stewardship-data-guidelines.pdf (accessed April 26, 2010).

Rosenzweig, Roy. 2003. "Scarcity or Abundance? Preserving the Past in a Digital Era." *American Historical Review* 108, no. 3 (June): 735–762. Available: www.historycooperative.org/journals/ahr/108.3/rosenzweig.html (accessed April 26, 2010).

Ross, Seamus. 2007. "Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries." Keynote address to the 11th European Conference on Research and Advanced Technology for Digital Libraries, Budapest, September 16–21, 2007. Available: www.ecdl2007.org/Keynote_ECDL2007_SROSS.pdf (accessed April 26, 2010).

Rusbridge, Chris. 2007. "Create, Curate, Re-Use: The Expanding Life Course of Digital Research." Paper presented at EDUCAUSE Australasia 2007. Available: hdl.handle.net/1842/1731 (accessed April 26, 2010). Used by permission of Chris Rusbridge, Digital Curation Centre.

UNESCO. 2003. *Guidelines for the Preservation of Digital Heritage.* Paris: Information Society Division, United Nations Educational, Scientific and Cultural Organization. Available: unesdoc.unesco.org/images/0013/001300/130071e.pdf (accessed April 26, 2010).

Wellcome Trust. "Open and Unrestricted Access to the Outputs of Published Research." London: Wellcome Trust. Available: www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Open-access/index.htm (accessed April 26, 2010).

Wikipedia. 2009. "Digital Dark Age." Wikipedia (March 5, 2010). Available: en.wikipedia.org/wiki/Digital_Dark_Age (accessed April 26, 2010).

The Changing Landscape

2

IN THIS CHAPTER:

As noted in Chapter 1, one of the factors that make digital curation necessary is the embedding of new ways of working in many scientific, scholarly, and research communities. These new ways of working are characterized by their reliance on networked computing and on the creation, management, use, and reuse of large data sets. Scholarship is already substantially data driven, and this will rapidly expand. The new ways of working are in turn influencing the management of digital information in libraries and archives.

The new data-driven scholarship has various terms associated with it, including cyberscholarship, e-science, e-research, derivative science, and cyberinfrastructure. The term *cyberscholarship* is more prevalent in the United States, whereas in other countries the terms *e-science* and *e-scholarship* are more commonly used. In this book *cyberscholarship* is the term used to refer to the ways in which networked computing, data, and scholars work together. The term *e-science* is used here to refer more specifically to research in scientific fields, and *cyberinfrastructure* is used to refer to what needs to be in place for the new ways of working.

This chapter examines in more detail the characteristics of the changing landscape, its requirements for digital curation, and the need for new kinds of skills to curate data. These topics are discussed in greater depth by Christine Borgman (2007) in *Scholarship in the Digital Age.*

Cyberscholarship: New Ways of Working

What, more precisely, are these new ways of working? What characterizes them?

Cyberscholarship is based on the availability of scholarly materials in digital form through computer networks. These scholarly materials range from large scientific data sets to digitized versions of the analog resources held in the collections of libraries and archives. (The nature of these scholarly materials is examined in more detail in Chapter 4.) New forms of research and scholarship are developing based on the availability of these digital materials and on computing techniques to analyze and present them; they differ significantly from traditional practices—Larsen (2008) notes "the transformative potential of digital scholarship." The new forms of research and scholarship include working in larger groups whose members may be based at different (perhaps many) geographic locations. The outcome is that research is collaborative, based on large digital data sets developed, shared, and used by international communities of scholars, although this is not yet widespread (Larsen, 2008).

Cyberscholarship is characterized not only by new collaborative structures but also by the enhanced ability to compute large quantities of data. New discoveries are made through these computations: more detailed analyses can be carried out and data can be visualized and simulated more readily. In the humanities, for example, new opportunities created by the conjunctions of data, networked computing, and scholars are encouraging new ways of carrying out scholarship. Examples from a 2008 symposium describe an art historian who reconstructs an ancient site in digital form, a professor of Romance languages who creates an interactive map illustrating how Spanish language and culture spread over time, and a linguist applying social network tools. The symposium's report notes that "Students and researchers alike are using simulation and interactive model-based learning. Mass digitization makes it possible to query large corpora of heterogeneous source materials, synthesize information across disciplines, and perform new types of analysis" (Smith, 2008: 1). Another example is the development of electronic cultural atlases. As well as including all of the features of their paper versions, such atlases also allow different research questions to be posed and new relationships to be identified by offering innovative ways of grouping, analyzing, and visualizing the data within them. They allow, for example, visualization through animated maps, enhanced search capabilities, and the engagement of a much wider community in providing the data. The Electronic Cultural Atlas Initiative (ecai.org) illustrates the possibilities.

Another characteristic of cyberscholarship is that it generates large quantities of data. This has major implications for how data are stored, managed, preserved over time, and used. To take one example: scholars can read and analyze only a small number of documents compared with the millions they can analyze using a high-speed computer. Computing of large data sets enables new kinds of results: "Profound research is possible by simple analysis of huge amounts of information. Computer programs can identify latent patterns of information or relationships that will never be found by human searching and browsing" (Arms, 2008).

Cyberscholarship places heavy emphasis on sharing and reusing data. Data sets may be unique and able to be reproduced only at great cost and sometimes not at all. Many consider that these data sets should be available for reuse by scholars other than those who developed and collected them. This implies the widespread adoption of standards for creating and collecting data to ensure they can be discovered, located, and preserved over time. A further implication is that ways of protecting the rights of data creators and owners are required.

A further characteristic of cyberscholarship is its reliance on the increasing availability of all kinds of materials in digital form, from digital versions of material created on paper and in other analog forms to the reports, drafts, data sets, images, video, and other materials that are increasingly created in digital form and exist *only* in digital form. This demands and drives the availability of these digital materials. It also suggests both that the roles of established institutions such as libraries and archives need to be redefined and that new kinds of institutions need to be developed to support scholarship.

Cyberscholarship in Practice

More examples of cyberscholarship in practice illustrate these characteristics. The National Virtual Observatory (us-vo.org) provides the means for researchers to group together data from astronomical data sets that are widely dispersed, allowing analysis of combined data sets to derive previously unattainable results. To provide this, the National Virtual Observatory has developed software such as an application programming interface (API), standards such as an XML encoding scheme for astronomical data, and applications that work with the API (Arms, 2008).

An informative collection of cyberscholarship examples in the humanities, social sciences, and scientific/technical/medical subject areas in the United States is available in a 2008 study carried out for the Association of Research Libraries (Maron and Kirby Smith, 2008). Some of these are "quite novel, making use of the space, speed, and interactivity that the Internet allows" (Maron and Kirby Smith, 2008: 9). One example noted is eBird (ebird.org/content/ebird), "harnessing the power of users" to develop a large central database of bird sightings submitted by amateur bird-watchers. The database is used by professional ornithologists and environmentalists (Maron and Kirby Smith, 2008: 27).

Visualization possibilities are demonstrated in the "Comparing Victorian & Second Life Immersive Environments" project (sydenham crystalpalace.wordpress.com). This presents in Second Life a virtual three-dimensional model of the Pompeii Court of the Sydenham Crystal Palace, which opened as a museum in South London in 1854.

Data-intensive websites that provide access to data sets and to the tools to extract data from them, analyze the data, and visualize them are increasingly common. An example is Data.gov, which aims to "increase public access to high value, machine readable datasets generated by the Executive Branch of the Federal Government" in the United States by providing easy access to these data sets and tools.

E-science

Much of the literature about digital curation comes from the e-science context. E-science is the term used in the United Kingdom to denote "the systematic development of research methods that exploit advanced computational thinking," enabling "new research by giving researchers access to resources held on widely-dispersed computers as though they were on their own desktops. The resources can include data collections, very large-scale computing resources, scientific instruments and high performance visualization" (www.rcuk.ac.uk/escience/default.htm). E-science activities in the United Kingdom are funded by Research Councils UK. Its characteristics are noted by British computer scientist David de Roure (Rusbridge, 2008a).

One characteristic is the increasing scale and diversity of participation; more people, both amateurs and professionals, can potentially participate. (This characteristic is capitalized on in collaborative sites such as eBird, noted earlier.) Another characteristic relates to the data, which, like participation, are also increasing in scale and diversity. New data collection tools and methods allow massive quantities of data that are increasingly more complex to be collected. Data are being shared more frequently through new online mechanisms, such as social tools (wikis, blogs, Twitter). Data, research, and journals are becoming readily available through the Open Access movement. Also being shared are scientific tools, such as workflows (noted in Chapter 4), which contribute to research becoming more easily repeatable, reproducible, and reusable. Arguably the most significant characteristic of e-science noted by de Roure is that it "is now enabling researchers to do some completely new stuff! As the pieces become easy to use, researchers can bring them together in new ways and ask new questions" (cited in Rusbridge, 2008a).

Cyberscholarship's Requirements and Challenges

To implement fully the new opportunities that cyberscholarship's new ways of working allow, different kinds of systems and facilities, that is, *cyberinfrastructure*, are needed. These include computer networks, libraries and archives, online repositories, and much more. New skill sets are also required. The requirements and challenges of cyberscholarship can be considered using William Arms's useful categorization of them into content, tools and services, and expertise (Arms, 2008; Nelson, 2009).

Content

Cyberscholarship requires that data are available for use and reuse in the future. Access to data is required as the foundation of high-quality scholarship, that is, scholarship that is based on verifiable data and that builds on them to lead to new knowledge. Access to data sets is required for analysis and querying by scholars who may be located anywhere in the world. But the data are often widely scattered, never made available, poorly archived, or even destroyed. Even if locatable and available, their use may be restricted by intellectual property rights or privacy legislation.

Changing processes of scholarly communication are also a major factor that is altering scholarship and curation practice. Scholarly communication has been based on the self-contained nature of books, journals, and conference proceedings as the key outputs of scholarship and on publishing structures and reward systems that acknowledge this self-contained nature. The data sets on which the publications are based have low priority in this process and, Nelson (2009) tells us, "are treated as second-class artifacts." The processes to capture, store, and manage data sets over time are not supported well—to quote Nelson again: "While the scientific process is becoming more data-driven, the scholarly communication process, even though largely automated, continues much as it has for hundreds of years."

In addition, the nature and quantities of data and the methods of their collection are changing. Take just one example—social networking sites. The issues of whether and how to preserve blogs are the subject of attention that is resulting in the development of workable solutions (e.g., Maureen Pennock's presentation at the 2009 iPres conference; Pennock, 2009). Facebook, Flickr, YouTube, and Twitter are also becoming mechanisms for the collection and sharing of data that are likely to be of interest to scholars in the future, but at present we lack viable ways of preserving them. The rapid developments in social networking sites also create problems for their curation.

Tools and Services

Effective and easy-to-use tools and services for locating, managing, analyzing, visualizing, and storing data are required for cyberscholarship, but sufficient of them are not yet available. Cyberscholarship demands data that need to be curated, and curation tasks, such as refreshing, migrating to new formats, tracking changes to data and verifying their provenance, must be much more effectively handled on a large scale. Automation of services is the key. (Automation of procedures and tools are noted in Chapter 13.) In the humanities, the shortage of such tools is acknowledged as a challenge. Better tools to make scholarly resources interoperable so that they can be used in other work were noted in a 2008 symposium (Smith, 2008). Specifically identified were federated searching services, better tools for ontology development, further development of standards, such as Open GIS (www.opengeospatial.org/standards), and better tools to track contributions to collaborative works. Also noted was the need for new models for teaching humanities scholars about how to collaborate.

Expertise

Nelson (2009) poses this question: "Who will capture this data, and where will it live?" Not only are the nature and quantities of data changing in cyberscholarship, but the ways they are acquired and stored are also changing (Nelson, 2009). The expertise required to acquire and store large quantities of data is in very short supply and, as Arms (2008)

The Internet Archive, founded in 1996, is a nonprofit organization that aims to preserve born-digital materials. It initially collected and archived webpages but has expanded to include moving images, texts, audio, and software. The Internet Archive collaborates with major libraries, archives, and museums around the world to preserve a record for generations to come. It places particular emphasis on open and free access to information and makes its collections readily available. The Wayback Machine (available at www.archive.org) is a search tool that provides access to websites archived by the Internet Archive.

notes, is concentrated in the Internet Archive (www.archive.org) and in commercial organizations such as Google, Amazon, and Microsoft. The skills required for curation are noted later in this chapter.

The existing infrastructure, based as it is on print-focused scholarly communication processes and on libraries and archives developed in response to these processes, is not proving to be adequate to accommodate the new demands of cyberscholarship. What is needed is a new form of infrastructure to address the challenges. The shape of what is needed, the *cyberinfrastructure*, is being developed and implemented, albeit slowly, and the role of libraries and archives within it is being vigorously debated. A 2009 OCLC report (Palmer, Teffeau, and Pirrman, 2009) suggests that the most important role for cyberinfrastructure is "providing the collections and tools needed for producing new scholarship." It notes that scholars are increasingly conducting their activities online, so the services that research libraries provide will need to be integrated into this digital work environment and that "good service will be defined by scholars' ability to find and use the digital information they need for all stages of research." The report concludes that the question for libraries and other institutions that provide information services is "not what services need to be offered digitally, but rather how do we proceed in the long term to move all services to an e-research platform" (Palmer, Teffeau, and Pirrman, 2009: 34).

In a traditional library, the user personally selects information by searching catalogs and browsing collections to locate and examine specific items. In new digital ways of working, large digital collections are searched and examined by computer programs directed by the user. For this to happen, digital content must be organized so computers can analyze it. This requires standardized data formats and software, as well as the ability to access data without legal barriers (Larsen, 2008). Curation and preservation of digital materials pose additional challenges, investigated in the chapters that follow.

The current laissez-faire approach to developing cyberinfrastructure is widely considered to be inadequate. A workshop jointly sponsored by the United States' National Science Foundation (NSF) and the United Kingdom's JISC proposed as a goal that, by 2015, "all publicly-funded research products and primary resources will be readily available, accessible, and usable via common infrastructure and tools through space, time, and across disciplines, stages of research, and modes of human expression" (Larsen, 2008).

PARSE.Insight (www.parse-insight.eu) is a research project funded by the European Union. (PARSE stands for Permanent Access to the Records of Science in Europe.) It has developed a roadmap to guide the development of a cyberinfrastructure for scientific data in Europe. This roadmap specifies "Organisational and Social Infrastructure concepts and components" that include policies to mandate the deposit of research data; robust and reliable places to deposit those data; and making publication of data "as valued and as referencable as is a publication of a paper in a journal" (PARSE.Insight, 2009: 12). The roadmap also notes "Technical Science Data Concepts and Components" that need to be addressed: create and maintain representation information; sharing of information about hardware and software; authenticity of a digital object; digital rights; persistent identifiers; transfer of custody and brokering services; and certified repositories.

New structures not yet envisaged will also develop. We could probably not have foreseen a few years ago the widespread implementation of cloud computing. It is now viewed as offering new possibilities: Chris Rusbridge (2008b: 217) conjectures "Can we combine the institution and the discipline to achieve network effects with institution components?", for example, the cloud's "mass appeal [and] highly scalable centralized services" with libraries?

Digital Curation: A New Profession, New Requirements

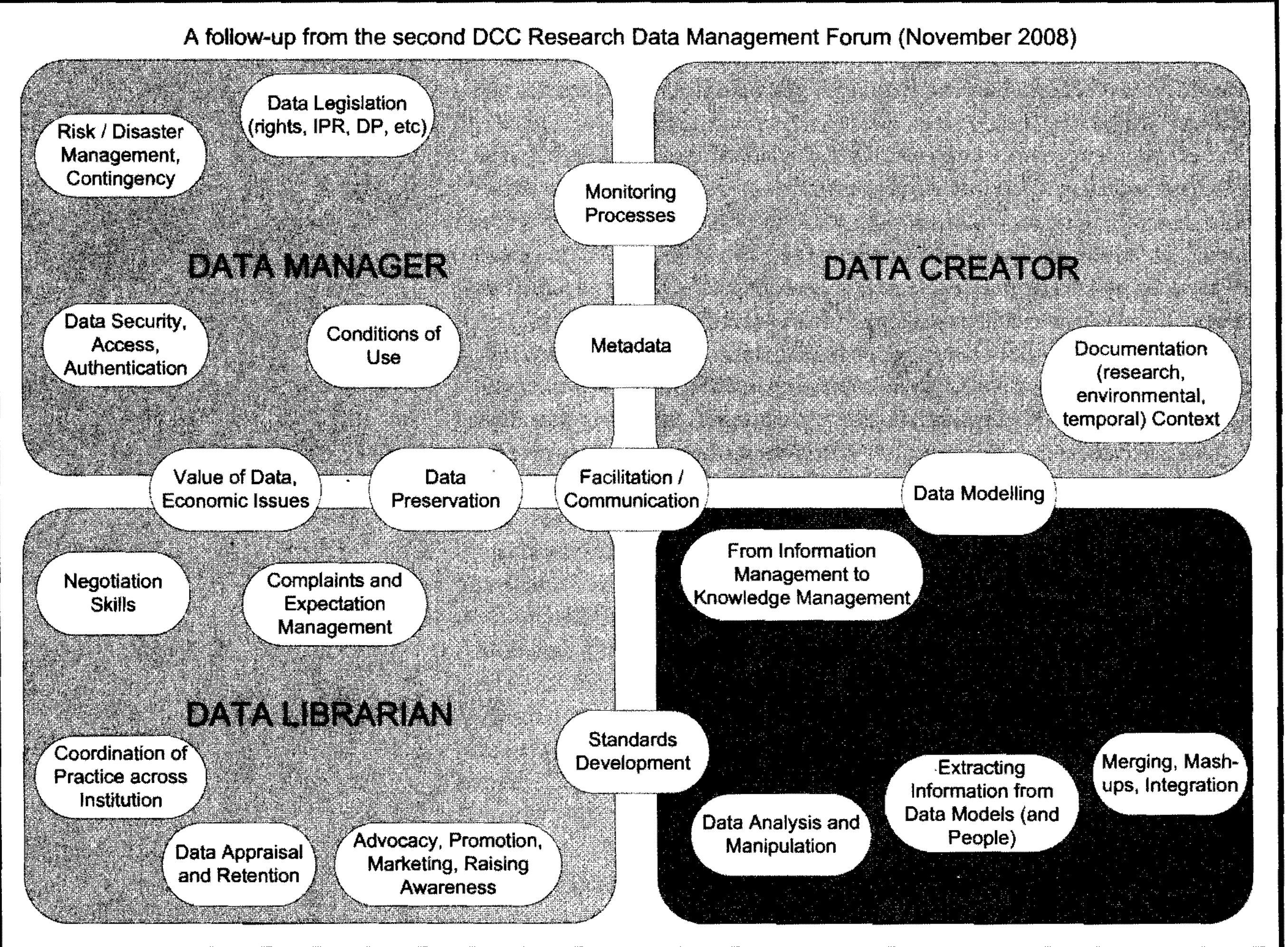
New skill sets are required for effective digital curation, and considerable research has been undertaken to identify and map these skill sets. Although this research has been carried out in different contexts, the results are similar.

One of the more comprehensive listings of skills is an outcome of the SHERPA (Securing a Hybrid Environment for Research Preservation and Access) Project (www.sherpa.ac.uk/index.html). The SHERPA Project is set in the context of institutional repositories and encompasses most aspects of curation as indicated in the Digital Curation Centre's Curation Lifecycle Model, although with different emphases (e.g., less importance is placed on use and reuse of data). The *Institutional Repositories: Staff and Skills Set* (Robinson, 2009) breaks down the knowledge and skills required by repository managers and administrators into nine categories: management; software; metadata; storage and preservation; content; advocacy, training, and support; liaison (internal); liaison (external); and current awareness and professional development. These categories are further developed; for example, the software category notes "familiarity with standard web-based software systems including (but not limited to) Unix, Linux, SQL Server, MySQL, SGML, XML, PHP, JAVA, PERL, Apache [and] at least one major repository software including (but not limited to) EPrints, DSpace, Fedora, OPUS." The specifics will change over time; what is more important and more useful is the identification of the areas where skills are required. These nine skills categories, identified in the institutional repository context, map well to other contexts.

In the archives context, Adrian Cunningham (2008: 541–542) proposes the "skills and capabilities required for digital archiving." In addition to a range of generic personal attributes, such as flexibility and research ability, the skills and capabilities he identifies for the archives domain can be mapped easily to SHERPA's areas. For example, Cunningham's "Auditing and compliance, QA, Preparing business cases, Modeling and analytical ability, [and] System design and implementation" fit comfortably in the management category identified by SHERPA. Both place strong emphasis on the areas of (to use SHERPA's terminology) Advocacy, Training and Support, and Liaison. Cunningham's phrase "out of the basement and into the boardroom" summarizes neatly one of the requirements, and his "Communication, influence, change management" and "Consultation and negotiation" terms also map closely.

In the e-science context, Pryor and Donnelly (2009) identify core skills for research data management. A model developed for the Research Data Management Forum (see Figure 2.1) indicates that different skills are needed by different players in research data management: data managers; data creators; data librarians; and data scientists ("RDMF2: Core Skills Diagram," 2008). Their list of key skills acknowledges data analysis and reuse ("Data Analysis and Manipulation"; "Extracting Information from

Figure 2.1. Core Skills for Data Management


A follow-up from the second DCC Research Data Management Forum (November 2008)
Risk / Disaster Management, Contingency
Data Legislation (rights, IPR, DP, etc)
Monitoring Processes
DATA MANAGER
DATA CREATOR
Data Security, Access, Authentication
Conditions of Use
Metadata
Documentation (research, environmental, temporal) Context
Value of Data, Economic Issues
Data Preservation
Facilitation / Communication
Data Modelling
From Information Management to Knowledge Management
Negotiation Skills
Complaints and Expectation Management
DATA LIBRARIAN
Standards Development
Coordination of Practice across Institution
Extracting Information from Data Models (and People)
Merging, Mash-ups, Integration
Data Analysis and Manipulation
Data Appraisal and Retention
Advocacy, Promotion, Marketing, Raising Awareness

Source: data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html. © Chris Rusbridge/Martin Donnelly/Research Data Management Forum. Developed for a meeting of the Research Data Management Forum in November 2008, and reproduced by permission of Martin Donnelly, Chris Rusbridge, and the Research Data Management Forum.

Data Models [and People]"; "Merging, Mashups, Integration") in a way that is not evident in the SHERPA list, but in other areas, for instance, "Facilitation/Communication" and "Negotiation Skills," it is in general accord with the SHERPA list.

Fyffe, Ludwig, and Warner's (2005: 4) *Sample Digital Preservation Curriculum Framework* is specifically focused on universities. Its five broad areas are General Awareness; Information Life Cycle Management (ILM); Information Storage Management and System Maintenance; Best Practices and Standards; and Legal Issues and University Policies. This list is also in broad agreement with the previous lists in that it does not focus exclusively, or even heavily, on the technical system and preservation aspects but emphasizes the context in which curation happens and the need for strong advocacy.

The most comprehensive listing of the knowledge and competencies is that developed by the DigCCurr Project based at the School of Information and Library Science, University of North Carolina at Chapel Hill. This project has among its outcomes two comprehensive listings, of "High-Level Categories of Digital Curation Functions" (Lee, 2008) and a "Matrix of Digital Curation Knowledge and Competencies" (Lee, 2009).

The DigCCurr Project's high-level categories can be mapped closely to the SHERPA Project's skills, as the partial comparison in Figure 2.2 indicates. The DigCCurr and SHERPA listings diverge when the use and reuse aspects of curation are considered. As already indicated these aspects

Figure 2.2. Comparison of Skills Required for Digital Curation

SHERPA Project	DigCCurr Project
Management	Management Administration Preservation Planning and Implementation Purchasing and Managing Licenses to Resources Analysis and Documentation of Curation Functions Evaluation and Audit of Curation Functions
Metadata	Description, Organization, and Intellectual Control
Storage and Preservation	Archival Storage Preservation Planning and Implementation Data Management
Content	Selection, Appraisal, and Disposition Destruction and Removal Analysis and Characterization of Digital Objects/Packages Validation and Quality Control of Digital Objects/Packages
Advocacy, Training, and Support	Advocacy and Outreach Education and Sharing of Expertise or Guidance on Curation Functions
Liaison (External)	Collaboration, Coordination, and Contracting with External Actors

are not emphasized by the SHERPA listing, which was developed in the context of institutional repositories. DigCCurr lists, for example, "Use, Reuse and Adding Value to Accessed Information" and "Reference and User Support Services."

Educating and Training Digital Curators

The recognition that new skills are required for digital curation has led to the development of courses through which students can learn curation skills. New courses of formal study are now emerging. These range from single subjects to full postgraduate programs at the master's level. Pryor and Donnelly (2009: 161–163) list some of the opportunities for acquiring the new skills required for digital curation. They point to the range of skills that are required, from "the basic toolkit that will equip researchers with the ability to plan for the longevity of their data to the more sophisticated suite of tools required by a data management professional that may need to be assimilated" (Pryor and Donnelly, 2009: 159).

Among the full academic programs at the postgraduate certificate or master's level is the MA in Digital Asset Management at King's College London. Its focus is on "curatorial and technical standards that arise throughout the 'digital resource life-cycle', from creation through management, access and dissemination to long-term preservation" (www.kcl.ac.uk/schools/humanities/depts/cch/pg/madam). Another program is the MSc in Information Management and Preservation offered by the Humanities Advanced Technology & Information Institute at the University of Glasgow (www.gla.ac.uk/departments/hatii), which aims to equip its graduates to work as archivists, records managers, and digital curators through developing their skills in digital curation and preservation issues as well as in the competencies required by these professions.

Other programs are specializations or concentrations within an established master's program, particularly within Master in Library and Information Science programs in the United States. An example of this is the Specialization in Data Curation within the Master of Science at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign. This focuses on "data collection and management, knowledge representation, digital preservation and archiving, data standards, and policy" in the research context (www.lis.illinois.edu/programs/ms/data_curation.html). Flexible delivery of some programs is also developing. The University of Arizona's fully online Graduate Certificate in Digital Information Management (digin.arizona.edu) combines "intensive, hands-on technology learning" with theoretical principles to develop the skills needed to manage large digital collections, including digital curation skills.

The importance of developing skills and knowledge in digital curation is acknowledged by scholarship and fellowship programs, such as those funded by the Institute of Museum and Library Services in the United

States. These include scholarships for students enrolled in the University of Arizona's Graduate Certificate in Digital Information Management and in the Carolina Digital Curation Doctoral Fellowship Program associated with the DigCCurr Project at the University of North Carolina at Chapel Hill.

Digital curation is a field where up-to-date skills and knowledge have typically been acquired on the job, supplemented by attendance at workshops and short courses. A wide range of workshops and courses are offered on an ongoing basis. One listing of these is available at the DCC's website (www.dcc.ac.uk). Summer schools and professional institutes are also available, such as DigCCurr's Curation Practices for the Digital Object Lifecycle (www.ils.unc.edu/digccurr/institute.html). Symposia in the field are well established, most notably the International Digital Curation Conference, iPres (rdd.sub.uni-goettingen.de/conferences/ipres/ipres-en.html) and the DigCCurr conference.

Training of researchers in digital curation awareness and skills is also receiving attention in many countries. In the Australian context, for example, where cyberinfrastructure is developing rapidly, work is taking place to determine the skills required. These need to be available through a range of venues and modes. While technical skills are often closely linked to specific disciplines, it is generally agreed that nontechnical skills are equally important. Generic skills such as project management, negotiation, team building, and problem solving were identified as important nontechnical skills. In addition, researchers need a general understanding of the e-research environment and how it can assist their research (Henty, 2008).

Summary: Meeting the New Demands

The rapid evolution of new kinds of scholarship based on the use and reuse of data is changing how many disciplines operate. The nature of participation in scholarship is also rapidly changing so that individuals can and do contribute through "citizen science" projects and online social networking mechanisms. To meet the challenges posed by this rapid evolution, changes in the kinds of infrastructure that support scholarship are essential and new skills are required. These will develop over coming years.

The next chapter introduces the DCC Curation Lifecycle Model, which provides a structure for the rest of this book. It also describes the OAIS (Open Archival Information System) Reference Model, a key conceptual model for digital curation.

References

Arms, William Y. 2008. "Cyberscholarship: High Performance Computing Meets Digital Libraries." *Journal of Electronic Publishing* 11, no. 1. Available: dx.doi.org/10.3998/3336451.0011.103 (accessed April 26, 2010).

Borgman, Christine L. 2007. *Scholarship in the Digital Age*. Cambridge, MA: MIT Press.

Cunningham, Adrian. 2008. "Digital Curation/Digital Archiving: A View from the National Archives of Australia." *American Archivist* 71 (Fall/Winter): 530–543.

Fyffe, Richard, Deborah Ludwig, and Beth Forrest Warner. 2005. "Digital Preservation in Action: Toward a Campus-Wide Program." *EDUCAUSE Center for Applied Research Research Bulletin* 19 (September). Available: www.educause.edu/ECAR/DigitalPreservationinActionTow/157552 (accessed April 26, 2010).

Henty, Margaret. 2008. "Developing the Capability and Skills to Support eResearch." *Ariadne* 55 (April). Available: www.ariadne.ac.uk/issue55/henty (accessed April 26, 2010).

Larsen, Ronald L. 2008. "On the Threshold of Cyberscholarship." *Journal of Electronic Publishing* 11, no. 1. Available: dx.doi.org/10.3998/3336451.0011.102 (accessed April 26, 2010).

Lee, Christopher A. 2008. "High-level Categories of Digital Curation Functions. Draft, Version 14." Chapel Hill, NC: DigCCurr (September 8, 2008). Available ils.unc.edu/digccurr/digccurr-funct-categories.pdf (accessed April 26, 2010).

———. 2009. "Matrix of Digital Curation Knowledge and Competencies (Overview). Draft, Version 13." Chapel Hill, NC: DigCCurr (June 17, 2009). Available: ils.unc.edu/digccurr/digccurr-matrix.html (accessed April 26, 2010).

Maron, Nancy L., and K. Kirby Smith. 2008. *Current Models of Digital Scholarly Communication: Results of an Investigation Conducted by Ithaka for the Association of Research Libraries.* Washington, DC: Association of Research Libraries. Available: www.arl.org/bm~doc/current-models-report.pdf (accessed April 26, 2010).

Nelson, Michael L. 2009. "Data-driven Science: A New Paradigm?" *Educause Review* 44, no. 4 (July/August): 6–7. Available: www.educause.edu/EDUCAUSE+Review/EDUCAUSEReviewMagazineVolume44/DataDrivenScienceANewParadigm/174196 (accessed April 26, 2010).

Palmer, Carole L., Lauren C. Teffeau, and Carrie C. Pirrman. 2009. *Scholarly Information Practices in the Online Environment: Themes from the Literature and Implications for Library Service Development.* Dublin, OH: OCLC. Available: www.oclc.org/programs/publications/reports/2009-02.pdf (accessed April 26, 2010).

PARSE.Insight. 2009. *Draft Road Map.* Didcot. Available: www.parse-insight.eu/publications.php#d2-1 (accessed April 26, 2010).

Pennock, Maureen. 2009. "ArchivePress: A Really Simple Solution to Archiving Blog Content." Paper presented at iPres (October 6, 2009). Available: www.cdlib.org/iPres/presentations/Pennockm.pdf (accessed April 26, 2010).

Pryor, Graham, and Martin Donnelly. 2009. "Skilling Up to Do Data: Whose Role, Whose Responsibility, Whose Career?" *International Journal of Digital Curation* 4, no. 2: 158–170. Available: www.ijdc.net/index.php/ijdc/article/viewFile/126/133 (accessed April 26, 2010).

"RDMF2: Core Skills Diagram." The RDMF Blog, comment posted December 17, 2008. Available: data-forum.blogspot.com/2008/12/rdmf2-core-skills-diagram.html (accessed April 26, 2010).

Robinson, Mary. 2009. *Institutional Repositories: Staff and Skills Set.* Nottingham: SHERPA. Available: www.sherpa.ac.uk/documents/Staff_and_Skills_Set_2009.pdf (accessed April 26, 2010).

Rusbridge, Chris. 2008a. "David de Roure on 'the new e-Science.'" Digital Curation Blog, comment posted September 15, 2008. Available: digitalcuration.blogspot.com/2008/09/david-de-roure-on-new-e-science.html (accessed April 26, 2010). Used by permission of Chris Rusbridge, Digital Curation Centre.

———. 2008b. "Tomorrow, Tomorrow, and Tomorrow: Poor Players on the Digital Curation Stage." In *Digital Convergence: Libraries of the Future* (pp. 207–217), edited by Rae Earnshaw and John Vince. London: Springer. Used by permission of Chris Rusbridge, Digital Curation Centre.

Smith, Kathlin. 2008. "Symposium Examines Research Topics at Nexus of Digital Humanities and Computing." *CLIR Issues* 65 (September/October). Available: www.clir.org/pubs/issues/issues65.html (accessed April 26, 2010).

Conceptual Models

3

This chapter describes and investigates the application of two key conceptual models for digital curation. The first, the Digital Curation Centre (DCC) Curation Lifecycle Model, outlines the actions that comprise digital curation and presents these actions in graphic form. This Curation Lifecycle Model is used as the structural basis of Parts II and III of this book. The second conceptual model, the Open Archival Information System (OAIS) Reference Model, is an International Organization for Standardization (ISO) standard that is widely used as the basis for the design and implementation of digital archival systems. Several other models of the lifecycle of data are also noted in this chapter.

An important difference between these two models must be noted. The OAIS Reference Model does not take account of activities outside the digital archival system: in particular, it does not offer guidance on the creation of data or on the use and reuse of data. The DCC Curation Lifecycle Model explicitly includes the activities that take place outside the archival system—that is, it describes *curation* rather than archiving or preservation alone.

Standards and models are significant in any context in which information is managed. This point is readily demonstrated. The National Information Standards Organization (NISO; www.niso.org/standards) publishes a considerable number of standards for library and information science, and there are many standards for archives management and preservation, for example, in the areas of archival description, preservation, and storage. A concise and informative summary of the use of standards in the library and information community, their importance, and their uses is provided by the National Library of Australia (2004).

For digital curation, standards and models are especially important. Although standards such as the OAIS Reference Model are applied widely, others need to be developed or be more widely adopted if digital curation challenges are to be more fully met. The need for standards is noted further in Chapter 8, and specific standards are referred to in following chapters.

IN THIS CHAPTER:

- ✓ The DCC Curation Lifecycle Model
- ✓ The OAIS Reference Model
- ✓ OAIS Functions
- ✓ OAIS and the DCC Curation Lifecycle Model
- ✓ Summary: The Importance of Models
- ✓ References

The DCC Curation Lifecycle Model

The DCC Curation Lifecycle Model is the structural basis of Chapters 5 to 15 in this book (Digital Curation Centre, 2008b). The DCC Curation Lifecycle is represented graphically on the back cover of this book. A draft version of the Model was published in 2007, and after a period of public consultation it was finalized in 2008. One example of its application is as the entry point to the listing of standards relating to digital curation and preservation (the DIFFUSE Standards Framework) on the DCC's website (Digital Curation Centre, 2008a).

A lifecycle model is particularly apposite to visualizing what happens when digital materials are curated. Actions applied (or not applied) at each stage of the information lifecycle directly influence how effectively that information can be managed and preserved in following stages of the lifecycle. For example, the addition of metadata during early stages of the lifecycle assists significantly in the long-term management of the data to which they are applied. Additionally, representing digital curation in a lifecycle model provides a checklist that can be used to ensure, when developing and implementing a curation plan, that all of the necessary stages are identified in the most appropriate order.

The DCC Curation Lifecycle Model offers a high-level overview of the activities that comprise digital curation. It is intended for organizations to use to model their digital curation activities, identifying the specific actions, technologies, standards, and skills required at each stage and adding to it or deleting from it where required. The Model was not designed for any specific digital curation operation or for application to any particular discipline. It can be applied in a wide range of digital curation contexts, including institutional repositories, digital archives, and electronic records management.

The DCC Curation Lifecycle Model notes three sets of actions: Full Lifecycle Actions, Sequential Actions, and Occasional Actions. The *Full Lifecycle Actions* are represented in the Model by four concentric inner rings: "Description and Representation Information," "Preservation Planning," "Community Watch and Participation," and "Curate and Preserve." These apply to every stage in the lifecycle. The innermost point (the bull's eye of the diagram) is "Data," indicating their centrality to the Model and emphasizing that it is data that are being curated.

The *Sequential Actions* in the outer ring represent the key actions needed to curate data as they move through their lifecycle, from their creation to their ultimate use and reuse. The sequence is not carried out once only; rather, it is repeated for as long as the data are being curated. This is indicated in the Model by the "Transform" action: data, through the process of being reused, can be transformed so that they form a new data set, which in turn needs to be created or received by an archive and so feeds back into the start of the lifecycle.

Occasional Actions may occur when specific conditions are met, but they do not apply to all data. For example, data may need to reappraised

CURATION LIFECYCLE MODEL

Full Lifecycle Actions

- Description and Representation Information
- Preservation Planning
- Community Watch and Participation
- Curate and Preserve

Sequential Actions

- Conceptualise
- Create or Receive
- Appraise and Select
- Ingest
- Preservation Action
- Store
- Access, Use, and Reuse
- Transform

Occasional Actions

- Dispose
- Reappraise

(hence the "Reappraise" action), or they may be disposed of as an outcome of the appraisal process (hence "Dispose").

The Full Lifecycle Actions are essential for the success of the curation process. They apply to most of the Sequential Actions. For example, "Preservation Planning" activities are ongoing activities that must be taken into account as data move through the Sequential Actions, and they are especially relevant to "Conceptualise," "Preservation Action," and "Access, Use, and Reuse." "Description and Representation Information" applies to all of the Sequential Actions—metadata (description information), for instance, is essential to all aspects of curation.

The DCC's website provides useful information about the DCC Curation Lifecycle Model (Digital Curation Centre, 2008c). It describes the Lifecycle's importance for data creators, data archivists, and data reusers. *Data creators* will find the Model relevant because the design of data has a crucial effect on their effective curation. Much of the information necessary for long-term curation and reuse needs to be captured when data are collected—for example, metadata that describes the data. The relevance of the Model to *data archivists* is that it identifies and describes what is required for effective data curation, assisting data archivists to ensure that the procedures and systems they develop are complete. *Reusers of other people's data* will find the Model relevant because access to data depends on how well they have been curated.

The DCC's website also indicates how the Curation Lifecycle Model can be applied in practice. It helps data curators ensure that their activities are appropriate by comparing these activities with the Lifecycle actions and developing actions that are missing or strengthening those that need it. The benefits of the Model are represented as enabling better mapping of activities against the Lifecycle; identifying weaknesses in practice; assisting in identifying collaborators (e.g., data creators) in the data curation process; supporting the documentation of policies and processes; encouraging the development of standards and technologies; and assisting with identifying tools and services for data curation. An extended digital curation lifecycle model based on the DCC's Model has been developed (Constantopoulos et al., 2009).

The Digital Curation Centre

The DCC is the organization that developed the DCC Curation Lifecycle Model, and it is also responsible for developing and promoting digital curation concepts and practices. It is a consortium of four major partners in the United Kingdom—the University of Edinburgh, the University of Glasgow, UKOLN (at the University of Bath), and the Science and Technology Facilities Council. It aims to support and promote continuing improvement in the quality of digital curation and digital preservation, particularly for the management of all research outputs in digital format. The DCC is a center of excellence in digital curation and preservation, providing authoritative expert advice and guidance on digital curation. Its website hosts a wide range of resources, software, tools, and support services. Although the DCC is based in the United Kingdom and has a primary audience in that country, its activities are highly significant for anyone working in the field of digital curation, no matter where they are based. In particular, the wide range of documentation about its activities and research it disseminates makes the DCC website an invaluable resource for anyone interested in digital curation.

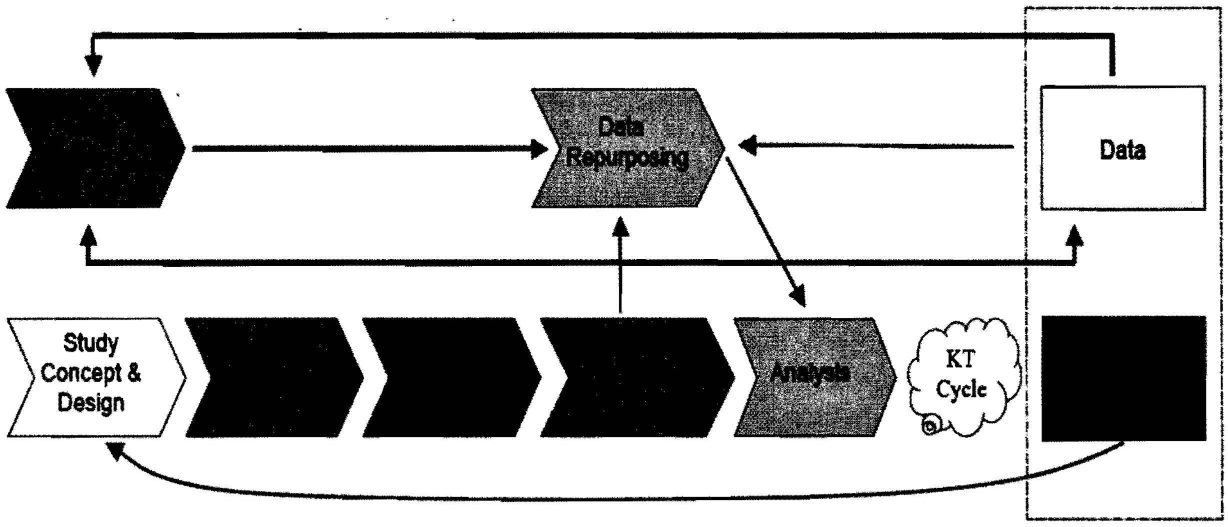
Other Lifecycle Models

The DCC Curation Lifecycle Model is only one of many models that presents what happens in digital curation. Three other models are noted here.

A simple four-step model is presented in a 2008 report on the stewardship of research data in Canada (Research Data Strategy Working Group, 2008). The four steps are Production, Dissemination, Long-Term Management, and Discovery and Repurposing. This model was developed for the specific purpose of assisting with the identification of gaps in the research data stewardship processes currently in place. It does not attempt to be comprehensive.

Humphrey provides a model of the lifecycle of research knowledge creation (see Figure 3.1). He notes that "Life cycle models are shaping the way we study digital information processes" (Humphrey, 2006: 1). They assist us to understand better the complex relationships that exist among the stages and activities in research. This improved understanding is essential for the development of useful data curation processes and practices that reflect actual practice and, therefore, stand a better chance of being widely adopted. As Humphrey (2006: 1) notes, "The life cycle approach makes us more aware of possible information losses in the gaps

Figure 3.1. The Life Cycle of Research Knowledge Creation



Source: Humphrey, 2006: 5. Reproduced by permission of Charles Humphrey.
Key: KT Cycle = Knowledge Transfer Cycle

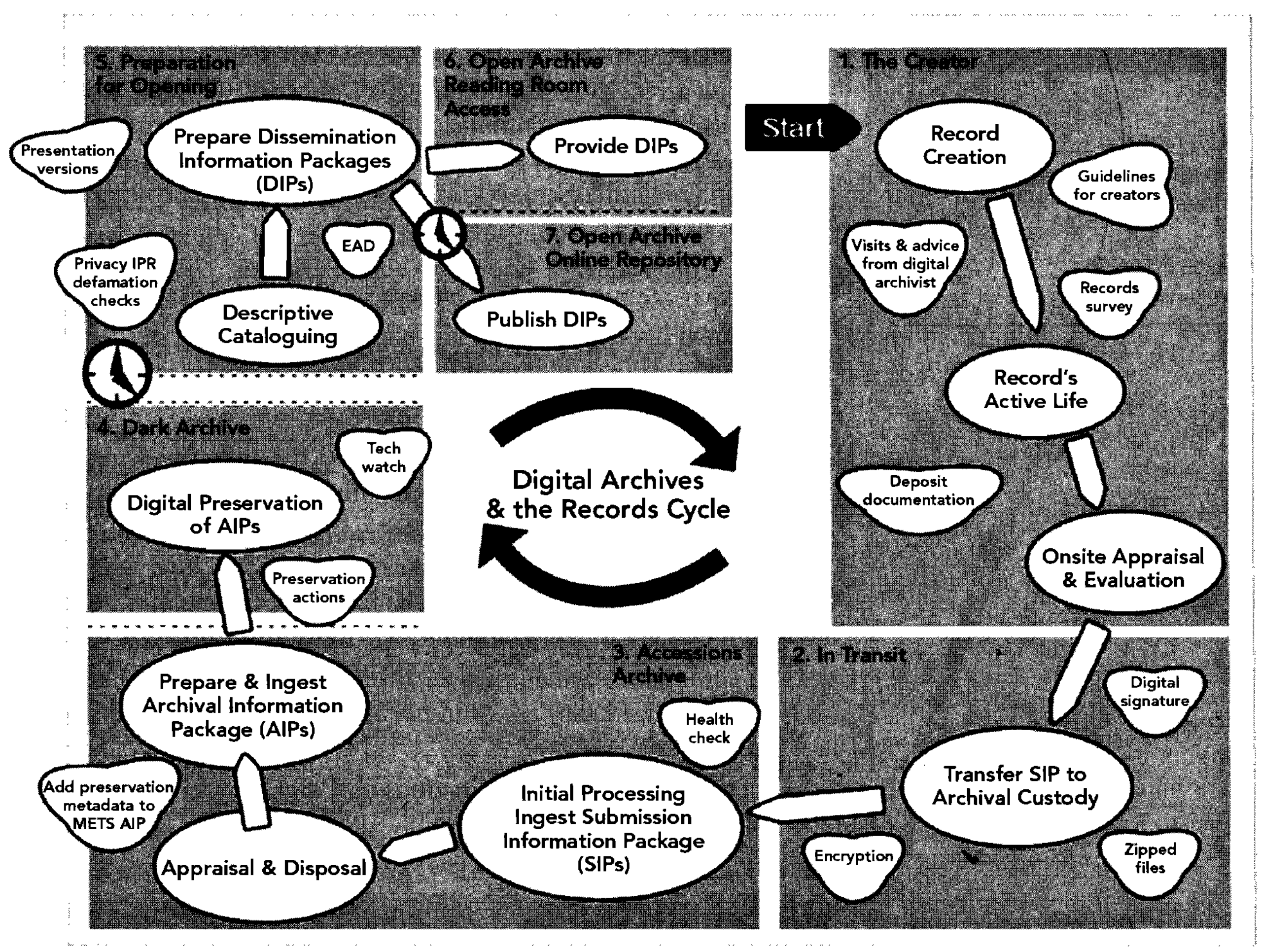
between stages." Note the strong emphasis on data in this model: data-driven research (e-science) focuses on the data, rather than the results, with reuse of data as a key component.

A more complex model comes from the Paradigm (Personal Archives Accessible in Digital Media) Project *Workbook on Private Digital Papers* (Paradigm Project, 2005–2007b). This model (see Figure 3.2) describes the long-term preservation of digital archives.

Many other lifecycle models are also relevant to digital curation, such as SHERPA's *A Lifecycle Model for an E-print in the Institutional Repository* (Knight, 2006). These have been developed for specific categories of material, whereas the DCC Curation Lifecycle Model is intended to be generic. They are not noted further in this book.

Higgins (2007) notes more lifecycle models. The significance of these lifecycle models, and of the DCC Curation Lifecycle Model and the OAIS Reference Model in particular, is that by mapping out the steps

Figure 3.2. Digital Archives and the Records Cycle


5. Preparation for Opening
Prepare Dissemination Information Packages (DIPs)
Presentation versions
EAD
Privacy IPR defamation checks
Descriptive Cataloguing
6. Open Archive Reading Room Access
Provide DIPs
7. Open Archive Online Repository
Publish DIPs
Start
1. The Creator
Record Creation
Guidelines for creators
Visits & advice from digital archivist
Records survey
Record's Active Life
Deposit documentation
Onsite Appraisal & Evaluation
Digital Archives & the Records Cycle
4. Dark Archive
Tech watch
Digital Preservation of AIPs
Preservation actions
3. Accessions Archive
Prepare & Ingest Archival Information Package (AIPs)
Health check
Add preservation metadata to METS AIP
Initial Processing Ingest Submission Information Package (SIPs)
Appraisal & Disposal
2. In Transit
Digital signature
Transfer SIP to Archival Custody
Encryption
Zipped files

Source: Paradigm Project, 2005–2007a. Reproduced by permission of the Bodleian Library, University of Oxford.

and actions needed for each specific context they describe, they allow comprehensive strategies and actions for effective digital curation to be planned and developed.

The OAIS Reference Model

The OAIS Reference Model is a widely adopted key standard for managing digital materials in a digital archiving system. It provides a generic framework for building a digital archive and is applicable to most actions in the curation lifecycle. The OAIS Reference Model has three primary aims. First, it provides a vocabulary of concepts related to preservation that is understood and adopted by people from a wide range of backgrounds (librarians, archivists, cultural heritage professionals, information technology [IT] personnel, scientists, scholars). Next, it defines an information model. Third, it defines a functional model: that is, it describes the key functions needed in a digital archive, and it provides information about the kinds of activities undertaken by each function.

The OAIS Reference Model was originally developed for the space data community in the 1990s. Input was sought from other interested communities to ensure that its concepts and terminology were commonly understood across different domains. It was widely adopted as a de facto standard (Consultative Committee for Space Data Systems, 2002) and formally adopted as ISO standard 14721:2003 (International Organization for Standardization, 2003). In 2009 it was being reviewed as part of the ISO's normal five-year review process.

The OAIS Reference Model is used as the basis for planning digital archives that are sustainable. Characteristics that contribute to its value as a planning tool include:

- clear identification of the responsibilities and interactions of data creators, users, and data archivists;
- definition of the processes required for effective long-term preservation and access to digital objects;
- the common language it establishes: this facilitates communication among archivists, librarians, data creators, and data users (who potentially come from a wide range of disciplines), each of whom has a different terminology to describe curation actions;
- unambiguous articulation of a framework for a digital archive, which significantly assists in planning and successful implementation; and
- detailed models of the functions of a digital archive. (Based on Higgins, 2006)

OAIS REFERENCE MODEL: KEY FUNCTIONS

Ingest
Archival Storage
Data Management
Administration
Access
Preservation Planning
Common Services

OAIS Functions

The OAIS Reference Model defines an Open Archival Information System that provides long-term information preservation and access. This system is "An archive, consisting of an organization of people and systems, that has accepted the responsibility to preserve information and make it available for a Designated Community." An OAIS archive is different from other kinds of archives by virtue of its meeting "a set of responsibilities, as defined in [the OAIS Reference Model]" (Consultative Committee for Space Data Systems, 2002: 1-1). The "Open" in OAIS refers to the development of the Model in open forums, not to the notion that access to the archive developed according to its criteria is unrestricted. The key functions of an OAIS, as defined in the Model, are summarized here:

- The **Ingest** function—the process of accepting information provided by Producers. Ingest is "responsible for receiving information from producers and preparing it for storage and management within the archive."
- The **Archival** storage function ensures that archival context remains secure and is stored appropriately—it "handles the storage, maintenance and retrieval of the AIPs [Archival Information Packages] held by the archive."
- The **Data Management** function supports access and updates information—it "coordinates the Descriptive Information pertaining to the archive's AIPs, in addition to system information used in support of the archive's function."
- The **Administration** function manages day-to-day operations and coordinates other functions.
- The **Access** function is the interface with the Designated Community—it "helps consumers to identify and obtain descriptions of relevant information in the archive, and delivers information from the archive to consumers."
- The **Preservation Planning** function develops preservation strategies, undertakes technology watch, etc. (Lavoie, 2000)

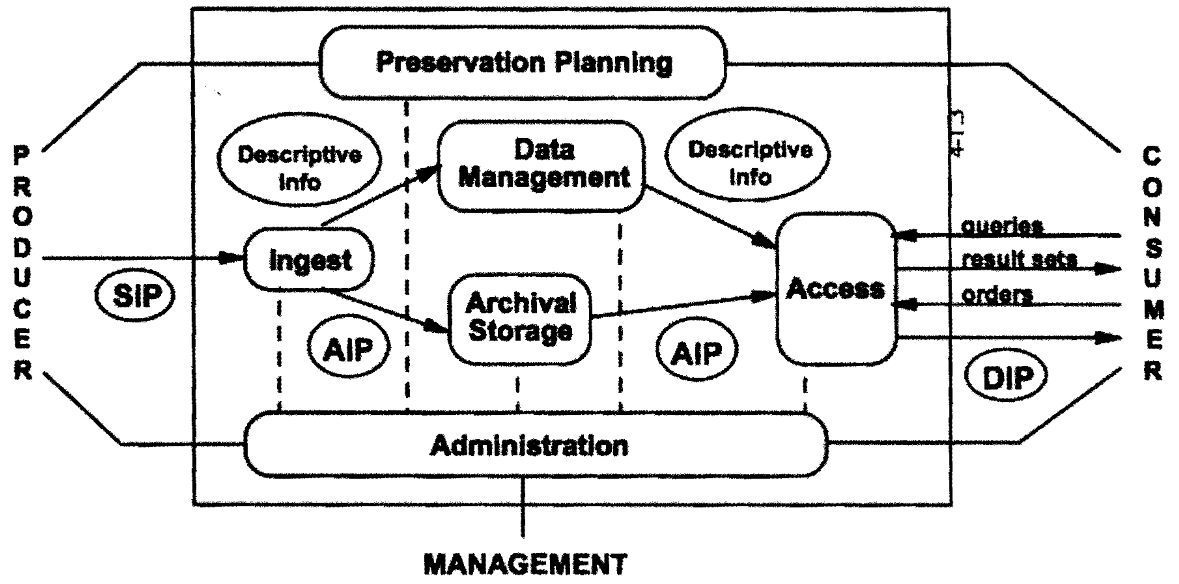
This is represented diagrammatically in Figure 3.3. A seventh function, **Common Services** (not noted in Figure 3.3), refers to the services that any IT system needs to function.

Actors and Objects

OAIS is based on the concept of actors and objects. *Actors* (who can be humans or computer systems) can perform in the roles of Producers, Managers, or Consumers. *Producers* are individuals, organizations, or computer systems that transfer digital information to the OAIS for preservation. *Managers* develop policy, define scope, and perform other management functions. *Consumers* are the individuals, organizations, or systems that are expected to use the information preserved by the OAIS.

An important concept in the OAIS Reference Model is the OAIS *Designated Community.* The Designated Community is a category of Consumer. It is the primary user group of the OAIS, to whom the OAIS must supply information that is understandable by this group. This means that the OAIS must have an understanding of the Designated

Figure 3.3. OAIS Functional Entities



Source: Consultative Committee for Space Data Systems, 2002: Figure 4.1.

Key: SIP = Submission Information Package
AIP = Archival Information Package
DIP = Dissemination Information Package

Community's knowledge base. Defining the knowledge base of the Designated Community requires data creators to think about who the users in the future might be and about the knowledge and understanding that future users can be assumed to have. Understanding potential future requirements allows data curators to better identify the data they need to preserve. *Objects* in OAIS are of three kinds: the Submission Information Package (SIP), the Archival Information Package (AIP), and the Dissemination Information Package (DIP).

OAIS Information Packages

OAIS is based on the concept of *information package*. An information package has three parts:

1. The digital object(s) to be preserved
2. The metadata required at that point in the system
3. Packaging information

The information package concept recognizes that a digital object consists of more than simply the digital content, in the form of a bit stream, which we want to preserve. An information package also includes information that we need in order to preserve the digital object, such as information about its attributes, or about what actions have been applied to it, and so on.

The OAIS Reference Model specifies three kinds of information packages. The *Submission Information Package* (SIP) is what arrives at the repository. It consists of the digital object, plus any descriptive and technical metadata accompanying the digital object and/or any other information the content provider considers relevant. SIPs may also be supplied to an OAIS from another digital repository.

The *Archival Information Package* (AIP) is produced by taking the SIP and adding to it, if required, further information about the digital object. The added information is either *Preservation Description Information* or *Representation Information*. Preservation Description Information is needed to manage the preservation of the digital objects submitted to the OAIS (referred to by the OAIS Reference Model as a *Content Data Object*). It has four components:

1. **Reference Information**: a unique and persistent identifier that assists in identifying and locating the Content Data Object
2. **Provenance Information**: the history of the archived Content Data Object
3. **Context Information**: information about the relationship of the Content Data Object to other objects; for example, the hierarchical structure of a digital archive
4. **Fixity Information**: a demonstration of authenticity, such as a hash value or checksum.

Representation Information is required to make the Content Data Object intelligible to its Designated Community. Representation Information is the technical metadata required to make the bit stream retrievable as a meaningful digital object. For example, a webpage that includes graphics needs descriptions of the web environment (browser, etc.), the text (ASCII standard), and the image files to make it understandable. (Representation Information is noted in more detail in Chapter 6.)

All of this (the Content Data Object, Preservation Description Information, and Representation Information) is parceled together as an AIP. The AIP is the basic element of the digital repository.

The DIP is produced when a user requests access to an object in the OAIS. The DIP consists of a copy of the Content Data Object plus any metadata and support systems necessary to retrieve and use the Content Data Object. The accompanying metadata and Representation Information will be a subset of all the metadata relating to that object. The nature of the metadata and Representation Information supplied is determined by the assumed knowledge of the Designated Community.

OAIS and the DCC Curation Lifecycle Model

The DCC's Curation Lifecycle Model follows the OAIS Reference Model closely. In particular, the sequential actions of the Lifecycle emulate what happens in OAIS. Figure 3.4 indicates the correlation (which, it should be noted, is not exact).

Figure 3.4. Correlations between the DCC Curation Lifecycle Model and the OAIS Reference Model

Curation Lifecycle Model	**OAIS Reference Model**	**Typical Actions** (RLG-OCLC, 2002)
Description and Representation Information	Relevant to all OAIS activities	
Conceptualise Create or Receive Appraise and Select Reappraise Dispose	Submission and "pre-ingest" activities (not part of the OAIS Reference Model)	• Check any existing deposit schedules to ensure everything expected has been received • Assign the digital object's unique identifier(s), if not already available, and provide labels for the physical artifact • Check for viruses and validate the integrity of the digital object and its physical carrier • Assess in detail the significant properties of the digital object, such as its look and feel, or functionality • Validate or improve the documentation • Where appropriate, reformat the digital object according to repository policies • Ensure that all necessary metadata for long-term maintenance and continuing access accompanies the object (Note: Some of these actions are also applied in later stages of the lifecycle)
Ingest	Ingest	• Assign and/or validate unique identifier • Select and validate the agreed underlying technology or underlying abstract form based on the object's significant properties • Transform the object as it was submitted, along with its associated metadata, into a byte stream that can be stored on suitable hardware in the repository • Establish necessary Representation Information • Verify all Preservation Description Information
Store Migrate	Archival Storage	• Move Archival Information Packages from Ingest into permanent storage • Manage the storage hierarchy • Refresh the storage media • Provide all necessary information to allow objects to be disseminated from the repository
Preservation Action	Data Management	• Develop pricing information (if applicable) and access controls • Develop customer profiles • Track user requests • Manage security information, including any usernames, passwords, digital certificates—anything used to authenticate users of the repository • Generate statistical information to improve operation
Preservation Planning Community Watch and Participation	Preservation Planning	• Monitor the designated community • Monitor technology • Monitor the significant properties of the repository's contents • Develop preservation strategies and standards for continuing access • Develop packaging designs and migration or routine transfer plans

(Continued)

Figure 3.4. Correlations between the DCC Curation Lifecycle Model and the OAIS Reference Model *(Continued)*

Curation Lifecycle Model	**OAIS Reference Model**	**Typical Actions** (RLG-OCLC, 2002)
Store	Administration	• Negotiate submissions agreements with content producers and providers • Review procedures • Maintain systems configurations for hardware and software • Develop and maintain repository policies and standards
Access, Use, and Reuse Transform	Access	• Prepare the Dissemination Information Package (DIP) • Verify the integrity of the information in the DIP • Ensure that users have permission for access to the material

Source: Consultative Committee for Space Data Systems, 2002; Digital Curation Centre, 2008b; RLG-OCLC, 2002.

Summary: The Importance of Models

Effective digital curation is based on the use of conceptual models. One model in particular, the OAIS Reference Model, is widely used as the basis for planning digital archives. It specifies the functions that a digital archive has to perform in order to preserve data and make it understandable to users over time. The DCC Curation Lifecycle Model (used as the structural basis of this book) is another model that provides guidance for planning and carrying out digital curation. The next chapter investigates in more detail what is meant by *data* and other related terms.

References

Constantopoulos, Panos, Costis Dallas, Ion Androutsopoulos, Stavros Angelis, Antonios Deligiannakis, Dimitris Gavrilis, Yannis Kotidis, and Christos Papatheodoro. 2009. "DCC&U: An Extended Digital Curation Lifecycle Model." *International Journal of Digital Curation* 4, no. 1: 34–45. Available: www.ijdc.net/index.php/ijdc/article/viewFile/100/75 (accessed April 26, 2010).

Consultative Committee for Space Data Systems. 2002. *Reference Model for an Open Archival Information System (OAIS): Recommendation for Space Data System Standards.* Washington, DC: CCSDS Secretariat. Available: public.ccsds.org/publications/archive/650x0b1.pdf (accessed April 26, 2010).

Digital Curation Centre. 2008a. "Browse All Standards by Lifecycle Action." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/resources/standards/diffuse/lifecycle/ (accessed April 26, 2010).

———. 2008b. *The DCC Curation Lifecycle Model.* Edinburgh: Digital Curation Centre, 2008. Available: www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed April 26, 2010).

———. 2008c. "Frequently Asked Questions about the DCC Curation Lifecycle Model." Edinburgh: Digital Curation Centre (July 2008). Available: www.dcc.ac.uk/digital-curation/digital-curation-faqs/dcc-curation-lifecycle-model (accessed April 26, 2010).

Higgins, Sarah. 2006. "Using OAIS for Digital Curation." Edinburgh: Digital Curation Centre (October 4, 2006). Available: www.dcc.ac.uk/resource/briefing-papers/introduction-curation/using-oais-curation (accessed April 26, 2010).

———. 2007. "Draft DCC Curation Lifecycle Model." *International Journal of Digital Curation* 2, no. 2 (December): 82–87. Available: www.ijdc.net/index.php/ijdc/article/viewFile/46/30 (accessed April 26, 2010).

Humphrey, Charles. 2006. "E-science and the Life Cycle of Research" (March 2006). Available: datalib.library.ualberta.ca/~humphrey/lifecycle-science 060308.doc (accessed April 26, 2010).

International Organization for Standardization. 2003. *Space Data and Information Transfer Systems—Open Archival Information System—Reference Model.* Standard 14721:2003. Geneva: International Organization for Standardization.

Knight, Gareth. 2006. *A Lifecycle Model for an E-print in the Institutional Repository.* Nottingham: SHERPA DP. Available: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.132.1916&rep=rep1&type=pdf (accessed April 26, 2010).

Lavoie, Brian. 2000. "Meeting the Challenges of Digital Preservation: The OAIS Reference Model." Dublin, OH: OCLC. Available: www.oclc.org/research/publications/archive/2000/lavoie/ (accessed April 26, 2010).

National Library of Australia. 2004. "National Library Standards Activities." Available: www.nla.gov.au/services/standards.html (accessed April 26, 2010).

Paradigm Project. 2005–2007a. "Digital Archives & the Records Cycle." In *Workbook on Digital Private Papers.* Paradigm Project. Available: www.paradigm.ac.uk/workbook/introduction/paradigm-lifecycle.html (accessed April 26, 2010).

———. 2005–2007b. *Workbook on Digital Private Papers.* Paradigm Project. Available: www.paradigm.ac.uk/workbook (accessed April 26, 2010).

Research Data Strategy Working Group. 2008. *Stewardship of Research Data in Canada: A Gap Analysis.* Ottawa: National Research Council Canada. Available: data-donnees.gc.ca/docs/GapAnalysis.pdf (accessed April 26, 2010).

RLG-OCLC. 2002. *Trusted Digital Repositories: Attributes and Responsibilities.* Mountain View, CA: Research Libraries Group. Available: www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf (accessed April 26, 2010).

Defining Data

4

This chapter investigates in more detail what is meant by the term *data* and by other related terms. The investigation of definitions is necessary because it allows an important question to be better addressed: What exactly is it that we want to curate?

Chapter 1 noted that the definitions of the terms *data*, *digital object*, and *database* (see sidebar) used in this book come from the DCC Curation Lifecycle Model (Digital Curation Centre, 2008). They are located in the innermost of the concentric rings (the bull's-eye) of the lifecycle and apply to all of the lifecycle's actions.

The precise meanings of these definitions need to be teased out. One way of doing this is to seek to answer further questions, such as these: Exactly what kinds of digital objects are the concern of digital curation? All kinds or just certain kinds? Exactly what kinds of databases?

Note that the term *data* as defined by the DCC and as used throughout this book is not limited to scientific data. Rather, it applies to any information in digital form, regardless of the context in which it is created, managed, and used. This could be in scientific and scholarly contexts but equally in libraries and archives, and it also applies to personal information in digital form.

As noted in Chapter 1, knowing the origin of the definitions used in the DCC Curation Lifecycle helps us understand the current concerns and emphases of digital curation better. Because of digital curation's origins in e-science there was initially a heavy focus on scientific data. This focus is now being widened to include humanities and social science data.

Data as Digital Heritage

UNESCO has a longstanding interest in the preservation of digital materials. Its *Charter on the Preservation of the Digital Heritage*, adopted in 2003, encompasses "information created digitally" as well as information "converted into digital form from existing analogue resources" and lists the types of digital materials as including "texts, databases, still and moving images, audio, graphics, software and web-

IN THIS CHAPTER:

Data refers to "any information in binary digital form." It includes *digital objects* and *databases.*

Digital objects can be simple or complex. "*Simple digital objects* are discrete digital items; such as textual files, images or sound files, along with their related identifiers and metadata. *Complex digital objects* are discrete digital objects, made by combining a number of other digital objects, such as Web sites."

Databases are "structured collections of records or data stored in a computer system."

(*Source*: Digital Curation Centre, 2008.)

pages, among a wide and growing range of formats" (UNESCO, 2003a: 1). This list is considerably expanded in UNESCO's *Guidelines for the Preservation of Digital Heritage.* In the *Guidelines* the materials that encompass digital heritage (carefully qualified with the words "at the time of writing"—this work was published in 2003) is wide-ranging and includes:

- **Electronic publications**: "information that is made available for wide readership," such as digital publications distributed via the web, or on carriers (CDs, DVDs, diskettes, e-book devices); they may be in traditional publication forms, such as monographs or serials, or there may be no analog equivalent, such as e-zines or websites.
- **"Semi-published" materials**: examples are pre-print papers and dissertations in e-print archives usually available for restricted use.
- **Organizational and personal records**: business and government records are increasingly created and managed in electronic records management systems.
- **Data sets**: data collected to "record and analyse scientific, geospatial, spatial, sociological, demographic, educational, health, environmental and other phenomena."
- **Learning objects**: developed for and used in educational settings, often in learning management systems.
- **Software tools**: software applications of all kinds.
- **Unique unpublished materials**: a very wide range of materials; research reports are an example.
- **Electronic "manuscripts"**: examples are drafts of works and personal correspondence.
- **Entertainment products**: produced by the film, music, broadcasting and games industries.
- **Digitally generated artworks and documentary photographs**.
- **Digital copies of nondigital materials**: examples of these include images, sound, text, and three-dimensional objects (derived from UNESCO, 2003b: 29–30).

Born-Digital *and* Digitized Data

Born-digital materials are defined by the Digital Preservation Coalition (2008) as materials "which are not intended to have an analogue equivalent, either as the originating source or as a result of conversion to analogue form." (This definition is, perhaps, too narrow. Born-digital materials, that is, materials created using a computer and therefore existing in a digital version, may also have an analog equivalent. The important point is the use of a computer to create the material.) *Digitized* materials are the result of a process of digitizing analog materials.

This distinction, between data that are born digital and data that are the product of a digitizing process, is not usually important for digital curation practice. Digital curation makes little distinction between born-digital and digitized data in most of the curation lifecycle actions—data, whatever their origin, still need to be appraised and selected, ingested, stored, and used and reused. A distinction is, however, made between born-digital and digitized data in the earlier actions of the lifecycle. The *Conceptualise* and *Create or Receive* Sequential Actions are intended to ensure that data are in the best possible shape to be curated; they attempt to ensure that data are preservation-ready. (Chapters 9 and 10 note this in more detail.)

The UNESCO *Guidelines* notes that the quantity of digital copies of analog materials is increasing rapidly but provides a caution—"Having originally been generated from non-digital sources, these might appear to be less vulnerable, but many of them are the only surviving version of originals that have since been damaged, lost or dispersed" (UNESCO, 2003b: 30). We often assume that an original analog version of material resulting from a digitizing project will remain in existence and such material is, therefore, less vulnerable than born-digital data because of the possibility of going back to the original analog version and digitizing it again. However, this assumption is not always correct. Examples abound of the disappearance of analog versions because they have been damaged, lost, deliberately destroyed, or dispersed. One example of such a loss was recently documented in accounts of an attempt to recreate a virtual version of the extensive collection of Islamic manuscripts in the Oriental Institute in Sarajevo after their destruction by Serb nationalists in 1992 (Riedlmayer, 2004). The high cost of digitizing may also be a major factor in determining whether it is possible to go back to an original analog version to digitize it again.

Data—And Much More

Another important consideration is that there is a wide variety of kinds of digital information and digital objects. The preceding example of digital heritage illustrates this point. The following examples of data generated by and used in e-science provide another illustration.

Discussions of digital curation often give the examples of data generated and used in science contexts. This is not surprising given digital curation's origins in e-science, as noted earlier in this chapter, and the increasing use of data in science. A key report from the National Science Foundation (2007: 21) illustrates the increasing use of data in relation to research and education for science and engineering, which are "increasingly data-intensive" because of networks, computers, and digital instrumentation. Each day immense quantities of data are produced, accessed, analyzed, integrated, and stored, as already noted in Chapter 1.

But the phrase "scientific data" used without qualification is not very helpful in getting a sense of its volume and diversity. In the scientific context, data are very diverse indeed. We can, however, identify some generic categories. One taxonomy is:

- **research collections**: for example, local data generated in a laboratory or research project;
- **community collections**: for example, genome databases such as MGI-Mouse Genome Informatics, "the international database resource for the laboratory mouse, providing integrated genetic, genomic, and biological data to facilitate the study of human health and disease" (Jackson Laboratory, accessed 2010); and
- **reference collections**: for example, the Protein Data Bank, containing "information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies" (Research Collaboratory for Structural Bioinformatics, accessed 2010) and providing tools and resources. (From the National Science Foundation taxonomy, as summarized by Liz Lyon, 2007: 15)

Other taxonomies have been developed. One characterizes scientific data as *canonical* or *episodic*. Canonical data does not change; episodic data does (e.g., information about climates). Another considers data as *raw*, *processed*, and *derived* data and *metadata*; yet another considers data in terms of its type: *"omics"* (fields of study in biology, such as proteinomics and genomics), *observational*, *simulations*, *multimedia*, *surveys*, *performances*, *computational*, *software*, and so on (Lyon, 2007: 15). Taxonomies such as these illustrate the variety, range, and different purposes of data. Although these taxonomies have been developed for and applied in science contexts, they can also be applied in other fields.

It is not just the very large and numerous scientific data sets that we wish to curate. Data cannot always be readily separated from the tools that act on them, analyze them, and interpret and present them. (An example in common use is an online currency converter, which manages constantly changing data and presents an analysis of that data on the fly.) Lyon notes that "many layers of subsequent interpretation act upon and transform the data" (Lyon, 2007: 15). New scientific methods are developing to analyze, manipulate, and present these data. New tools are being developed to mine, analyze, and visualize large data sets: examples are DNA sequence analysis, sky surveys, and "monitoring socioeconomic dynamics over space and time" (National Science Foundation, 2007: 21). The National Science Foundation (2007: 22) definition of data encompasses all of these additional layers:

> any and all complex data entities from observations, experiments, simulations, models, and higher order assemblies, along with the associated documentation needed to describe and interpret the data.

Note the words "simulations, models, and higher order assemblies"; that is, the products of manipulating data sets are included in this definition. The definition also includes "the associated documentation needed to describe and interpret the data"; this is noted in detail in Chapter 6. The acceptance of this definition expands the range of data and digital objects we may wish to curate.

Computer simulations are an example of these newer kinds of digital objects. These are computer programs that imitate or represent how something works and are used for purposes such as observing how systems might behave or to find out how a system operates. They are in common use in many scientific contexts. nanoHUB, a web resource for nanotechnology, is one example. It "offers simulation tools which you can access from your web browser, so you can not only learn about but also simulate nanotechnology devices" (Network for Computational Nanotechnology, accessed 2010). Examples of simple online simulations are plentiful, such as a simulation of an early wooden printing press (atlas.lib.uiowa.edu/press-animation.html) and some simple simulations in learning objects developed for grades K–6 (www.eduplace.com/kids/hmsc/content/simulation/).

Visualizations are another example of these newer kinds of digital objects. They apply computer programs to data sets in order to communicate results, summaries, or other characteristics of the data set in a visual form. A simple example of visualization can be seen at the nanoHUB website (nanohub.org/about) where analysis of the usage of this resource over time is presented both on a map of the world and as a graph.

FURTHER VISUALIZATION EXAMPLES

- "Many Eyes: Listing Visualizations" (manyeyes.alphaworks.ibm.com/manyeyes/visualizations?sort=rating)
- "Flowing Data" (flowingdata.com/category/visualization)

Simulations and visualizations may require curation, and this poses many questions: Do we curate the result of the simulation or visualization, or the application and the data set it was applied to, or just the data set? Do we curate all of the results generated by running simulation or visualization applications? Some visualizations are very large files, so keeping them may not be feasible in terms of the resources that are needed for their curation and for their appraisal (see Chapter 11).

There are other kinds of digital objects to curate, too, such as those that handle the processes of searching, integrating, and analyzing scientific data. Workflow software is increasingly being adopted by scientists. It allows scientists to quickly put together a set of software applications that automates data-handling procedures. Not only do workflows speed up scientific processes by automating repetitive processes, they also provide a means of reliably replicating scientific practice. An example of workflow software is the open source Taverna workbench (taverna.sourceforge.net). This allows scientists using a standard desktop computer (PC, UNIX, or Apple) to develop workflow processes and run them from their desktop. Goble and de Roure (2008: 11) summarize the significance of workflow software as "a reliable and transparent means for encoding a scientific method that supports reproducible science and the sharing and replicating of best-of-practice and know-how through reuse." Workflows, like the data sets on which they operate, also require systematic curation.

Metadata Is Data Too

Metadata, according to the National Science Foundation (2007: 22), "summarize data content, context, structure, interrelationships, and provenance (information on history and origins). They add relevance and purpose to data, and enable the identification of similar data in different data collections." In Chapter 6, this list of functions that metadata perform is added to and described in detail. The key point to note at this point is that metadata, too, need curation. The National Science Foundation (2007: 22) also includes in its definition of data that requires curation "the associated documentation needed to describe and interpret the data." In the DCC's Curation Lifecycle this is called *Description Information* and may also need to be curated. Chapter 6 examines metadata in more detail.

Databases

Databases are specifically noted in the DCC Curation Lifecycle as "structured collections of records or data stored in a computer system" (Digital Curation Centre, 2008). Data in databases are structured and controlled by database management software. Databases are widely used in contexts where their preservation is mandatory for a range of reasons, one being compliance with government regulations. Their curation poses many problems, not the least of which is their often constantly changing content.

Curation of databases is carried out in various ways. One of these requires taking snapshots of the database at specified intervals of time and curating these snapshots separately from the database. Another approach is to remove selected records from the live database and curate these separately. This process places a heavy emphasis on appraisal and selection (see Chapter 11) of records in the database in accordance with specified requirements (Müller, 2009).

Summary: New Kinds of Data

We are now in a better position to answer the question posed at the start of this chapter—What exactly is it that we want to curate? As well as data, digital objects, and databases (as defined by the DCC Curation Lifecycle), metadata, documentation, and "higher order assemblies" such as the examples of visualization and workflows noted above need to be curated. This curation will need to take place in all contexts, ranging from the handling of personal digital information to large data sets in science and humanities contexts.

And there will be more. New forms of digital heritage and new uses for already existing types of data will emerge. (Consider how social networking sites such as Facebook and Twitter have altered in a very short time the ways in which people and groups communicate and exchange data). Procedures, practices, and theories of digital curation must be open and flexible enough to accommodate new kinds of data and digital objects and new ways of using them.

The next chapter examines what is involved in digital curation—who is responsible for it and who does it—and considers the different roles of digital curators, archivists, preservation administrators, and other players in the curation arena.

References

Digital Curation Centre. 2008. *The DCC Curation Lifecycle Model.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/docs/publications/DCCLifecycle.pdf (accessed April 26, 2010).

Digital Preservation Coalition. 2008. "Introduction." In *Preservation Management of Digital Materials: The Handbook.* York: Digital Preservation Coalition (November 2008). Available: www.dpconline.org/graphics/intro/definitions.html (accessed April 26, 2010).

Goble, Carole, and David de Roure. 2008. "Curating Scientific Web Services and Workflows." *EDUCAUSE Review* 43, no. 5: 10–11.

Jackson Laboratory. "About MGI." Bar Harbor, ME: Jackson Laboratory. Available: www.informatics.jax.org/mgihome/projects/aboutmgi.shtml (accessed April 26, 2010).

Lyon, Liz. 2007. *Dealing with Data: Roles, Rights, Responsibilities and Relationships Consultancy Report.* Bath, England: UKOLN. Available: www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.doc (accessed April 26, 2010).

Müller, Heiko. 2009. "Database Archiving." Edinburgh: Digital Curation Centre (February 2, 2009). Available: www.dcc.ac.uk/resources/briefing-papers/introduction-curation/database-archiving (accessed April 26, 2010).

National Science Foundation. 2007. *Cyberinfrastructure Vision for 21st Century Discovery.* Arlington, VA: National Science Foundation. Available: www.nsf.gov/pubs/2007/nsf0728/index.jsp (accessed April 26, 2010).

Network for Computational Nanotechnology. "nanoHUB.org: About Us." West Lafayette, IN: Network for Computational Nanotechnology. Available: nanohub.org/about (accessed April 26, 2010).

Research Collaboratory for Structural Bioinformatics. "PDB: An Information Portal to Biological Macromolecular Structures." Piscataway, NJ: Research Collaboratory for Structural Bioinformatics. Available: www.rcsb.org/pdb (accessed April 26, 2010).

Riedlmayer, A. 2004. "The Bosnian Manuscript Ingathering Project." In *Ottoman Bosnia: A History in Peril* (pp. 27–38), edited by M. Koller and K Karpat. Madison: University of Wisconsin Press.

UNESCO. 2003a. *Charter on the Preservation of Digital Heritage.* Paris: United Nations Educational, Scientific and Cultural Organization. Available: portal.unesco.org/ci/en/files/13367/10700115911Charter_en.pdf/Charter_en.pdf (accessed April 26, 2010).

———. 2003b. *Guidelines for the Preservation of Digital Heritage.* Paris: Information Society Division, United Nations Educational, Scientific and Cultural Organization. Available: unesdoc.unesco.org/images/0013/001300/130071e.pdf (accessed April 26, 2010). *Guidelines for the Preservation of Digital Heritage*, March 2003 (pp. 29–30), © UNESCO 2003; used by permission of UNESCO.