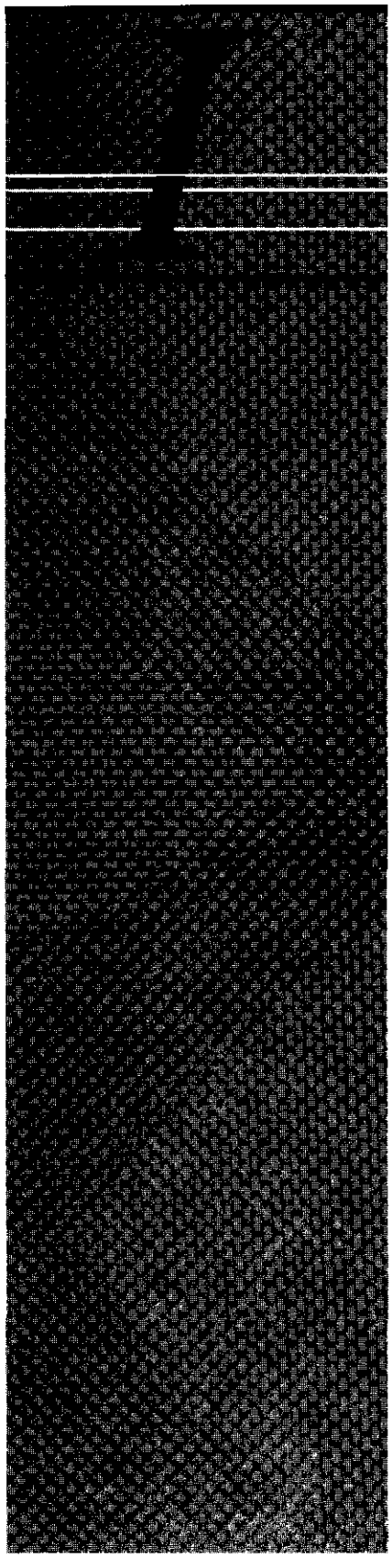
Digital Curation: Scope and Incentives



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.

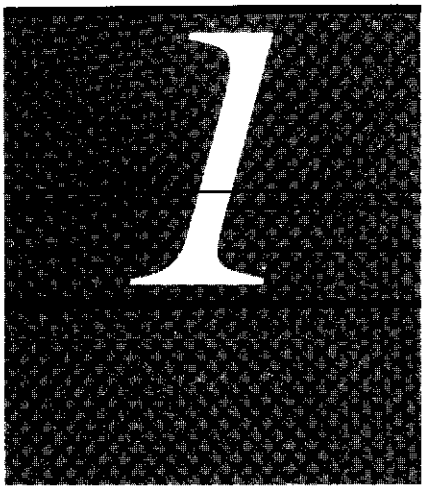
Chapter 1 indicates the reasons why digital curation is necessary, identifies what digital curation encompasses, suggests why you should be interested in digital curation, notes the main incentives for digital curation, and examines who does digital curation and what tasks they carry out.

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



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.

IN THIS CHAPTER:

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 docs 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/about1.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).