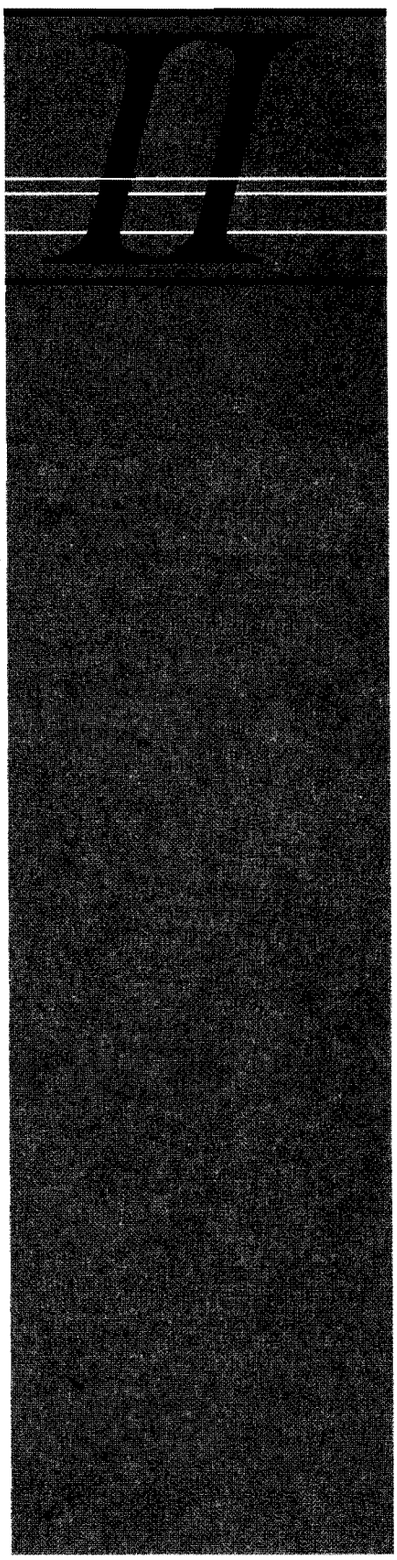
Key Requirements for Digital Curation



The four chapters in Part II examine the four Full Lifecycle Actions of the DCC Curation Lifecycle Model. These Full Lifecycle Actions define and describe the essential basic requirements for all aspects of digital curation. They apply to all of the Sequential Lifecycle Actions that are noted in Part III.

Chapter 5 covers the Full Lifecycle Action *Curate and Preserve*. It examines what digital curation aims to do, gives an overview of how these aims are achieved, and notes the roles of data creator, users, and curator.

Chapter 6 examines the Full Lifecycle Action *Description and Representation Information*—the metadata and other information that are essential for effective data curation.

Chapter 7 notes the essential nature of planning in data curation by describing another Full Lifecycle Action, *Preservation Planning*. It describes the need for planning at all stages of curation, notes the importance of developing policy for all aspects of digital curation, and refers to recent research into the costs of digital curation.

Chapter 8 completes the examination of Full Lifecycle Actions by describing the role of and key activities associated with *Community Watch and Participation*—the process of keeping up-to-date and participating in developments to advance and improve curation activities.

Curation and Curators

Chapter 5 examines the Digital Curation Centre (DCC) Curation Lifecycle's Full Lifecycle Action *Curate and Preserve*. The Full Lifecycle Actions, as noted in Chapter 3, apply to all stages of the Lifecycle and in the Lifecycle's diagram are surrounded by the Sequential Actions. They relate not just to the other Full Lifecycle Actions next to them in the Lifecycle diagram. They are also interdependent; for example, although Curate and Preserve is largely concerned with management and administration, it also applies to another Full Lifecycle Action, Description and Representation Information (see Chapter 6), which produces information that needs to be managed and administered. Conversely, Description and Representation Information is also required for effective management of data.

The activities that comprise the Curate and Preserve Full Lifecycle Action are stated in the DCC Curation Lifecycle Model as: "Be aware of, and undertake management and administrative actions planned to promote curation and preservation throughout the curation lifecycle" (Digital Curation Centre, 2008). This chapter examines in more detail what digital curation aims to do, what these management and administrative actions are, and who carries them out.

IN THIS CHAPTER:

Aims of Digital Curation

In Chapter 1 the differences between digital preservation and digital curation were noted. In this chapter these differences are revisited, and the aims of digital curation are examined further. *Digital preservation* refers to activities aimed at ensuring that we can access data, digital objects, and databases in the future—for longer than the life span of the software and hardware used in their creation and initial maintenance. Many definitions of digital preservation exist. Here are three of them:

1. "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).

2. "The managed activities necessary for ensuring both the long-term maintenance of a bytestream and continued accessibility of its contents" (RLG-OCLC, 2002: 3).
3. "The processes of maintaining accessibility of digital objects over time" (UNESCO, 2003: 161).

The key point of these definitions is that digital preservation is a set of managed activities aimed at ensuring that the bit stream is maintained and that data in all forms are accessible for a defined period of time. But digital preservation is not the same thing as digital curation. Chapter 1 noted the brief definition of ***digital curation*** provided by the DCC, which emphasizes adding value, trusted information, and active management of data in addition to preservation (Digital Curation Centre, "DCC Charter").

Digital preservation is a necessary part of curation, but by itself it is not sufficient. Preserving just the data, for example, copying the bit stream onto new forms of data storage, does not ensure that the digital objects that these data comprise will be usable in the future. This is where digital curation comes into play. It requires that data be actively managed and appraised so that their integrity is protected and their value is enhanced, with the aim of making them useful and usable in the future. To do this, we have to actively manage data over the whole of their life. This is where the Curate and Preserve action in the Curation Lifecycle comes into play. It requires us to know about, and apply, management and administrative actions that promote curation and preservation wherever they apply in the curation lifecycle.

Digital curation aims to produce and manage data in ways that ensure they retain three characteristics: longevity, integrity, and accessibility. *Longevity* refers to the availability of the data for as long as their current and future users (the Designated Community; see Chapter 6) require them. The life span of data is short unless action is taken. The length of time that data need to be maintained varies, but the minimum period of time usually exceeds the life expectancy of the access system (the hardware and software designed to view and/or use the data). *Integrity* refers to the authenticity of data —that they have not been manipulated, forged, or substituted. Because digital preservation techniques such as migration inevitably alter the data, authenticity has to be demonstrated by paying attention to such characteristics as provenance (where the data came from) and context (the circumstances surrounding the creation, receipt, storage, or use of data and their relationship to other data). *Accessibility* requires that we can locate and use the data in the future in a way that is acceptable to their designated community. For example, an image (such as a PDF) may be acceptable for some digital objects (such as documentation), but for other objects (a database, for example) the ability to manipulate or interrogate that object may be required by its Designated Community in the future.

Scope of Digital Curation

How are the aims of ensuring the integrity of digital objects over time and maintaining access to them achieved? To achieve the aims the digital

objects need to be managed, and this is the focus of the Curate and Preserve action described in this chapter. Some techniques that are commonly applied are listed here. (This section is based on the National Library of Australia's "Recommended Practices for Digital Preservation," accessed 2010.)

Ensuring Longevity

The main digital curation practices in common use that ensure the long life of digital objects include:

- refreshing data (moving data to a newer version of the same storage medium, or to different storage media, with no changes to the bit stream);
- checking accuracy of the results of refreshing;
- generating metadata that document the processes applied to refreshing data;
- maintaining multiple copies of the bit stream; and
- keeping track of changes (especially obsolescence) in hardware, software, file formats; and standards that might have an impact on digital preservation.

Ensuring Integrity

The main practices in common use that ensure the authenticity of digital objects include:

- refreshing data;
- checking accuracy of the results of refreshing;
- generating metadata that document the processes applied to refreshing data;
- protecting data by managing them in accordance with good IT practices for data security, backups, and error checking;
- maintaining multiple copies of the bit stream; and
- managing intellectual property and other rights.

Maintaining Accessibility

The key practices in current use include:

- maintaining the ability to locate digital materials reliably by assigning persistent identifiers to them to ensure they can be found;
- recording sufficient representation information for digital objects so that the bit stream is still meaningful and understandable in the future;
- producing digital objects in open, well-supported standard formats;
- limiting the range of preservation formats to be managed (often by normalizing data to standard formats);

- keeping track of changes (especially obsolescence) in hardware, software, file formats, and standards that might have an impact on digital preservation; and
- maintaining multiple copies of the bit stream.

There is considerable overlap among the three groups of techniques. Each of these practices can be linked to at least one, and usually more, of the specific actions in the Curation Lifecycle Model. For example, the practices of producing digital objects in open, well-supported standard formats and limiting the range of preservation formats to be managed are included in the Conceptualise and Create or Receive actions (see Chapters 9 and 10); and the practice of keeping track of developments that have an impact on digital preservation is part of the Community Watch and Participation action (see Chapter 8).

Roles of Digital Curators

Whose job is it to apply and manage the actions required for digital curation? What do those people do? Digital curation encompasses a wide range of tasks. The curation of scientific data is well documented. Some of the practical and technical curation tasks for scientists and research groups are:

- applying open-source software and open standards to encourage interoperability among different software and hardware platforms;
- creating metadata and annotations so that digital objects can be reused;
- linking related research materials and making sure the links are persistent;
- using persistent identifiers;
- being consistent about citation formats;
- deciding which digital objects need to be curated over the longer term;
- keeping data storage devices current; and
- validating and authenticating migrated data. (For a more detailed listing of these tasks, see Pennock, 2006.)

Several DCC case studies describe what curation involves in practice for some areas of science (Digital Curation Centre, "Case Studies"). One of these examines the curation of geospatial data (data that identifies the geographic location of features on the Earth's surface). According to the author of this case study (McGarva, 2006), curators of geospatial data:

- create data in a reliable manner that makes their context clear;
- implement sound creation practices to ensure that their data are reusable and sustainable over time;
- clearly identify any processing the data have undergone;
- establish standards for appraisal and selection of geospatial data;

- appraise and select geospatial data for long-term curation according to standards;
- work with data creators to ensure sufficient metadata is available; and
- maintain a technology watch to ensure that data formats do not become obsolete.

While the preceding curation tasks and activities are based on practice in science contexts, nearly all of them apply more widely. For example, metadata is of equal importance for long-term management of all kinds of digital objects in all contexts.

Digital curation also requires the sharing of responsibilities. It is important to determine who the stakeholders in digital curation are, because each kind of stakeholder is likely to have different knowledge and skills and different understandings about what the digital objects are and how they are used. Several kinds of stakeholders play differing roles in the data curation process: funding bodies, discipline-based groups (e.g., scientific organizations), data creators, data users and reusers, and data curators, each with different skills, understandings, and interests.

Funding Bodies

Funding bodies support data creation by providing money for research projects or digitization projects. These bodies are increasingly concerned with making sure that the data whose creation they fund are available to a wide range of users. More and more of them require grant applications made to them include provision for digital curation (as noted already in Chapter 1). Common examples of these requirements are a data management plan, or a plan for depositing data into a publicly accessible data repository. The process of developing and submitting grant applications, and reporting on progress, generates rich documentation about the purpose of data and their nature, along with other useful information. Documentation of this kind is crucial for the reuse of those data in the future.

Discipline Groups

Groups organized around a particular discipline or set of disciplines, particularly in the sciences, have a strong interest in digital curation. They may, for example, provide software that supports data handling. An example is the Protein Data Bank (www.rcsb.org/pdb), used by biologists in fields such as structural biology, biochemistry, genetics, and pharmacology, which has a "Software Tools" link to tools for data extraction and deposition preparation, format conversion, data validation, and dictionary and data management. Another example is ArchNet, which lists archaeology software tools (archnet.asu.edu/resources/Selected_Resources/Software/general.php). They may also establish and support data repositories, such as the Global Biodiversity Information Facility (www.gbif.org). Conway (2009) notes that stakeholders, while supporting digital curation as it applies to digital objects within their own field, are not usually concerned with other disciplines. She indicates a less positive

aspect of their role in digital curation, particularly in "emerging and specialist areas of scientific investigation, where much knowledge was still embedded in the data-using scientific groups, and there tended not to be mature documentation supporting the data" (Conway, 2009: 20–21).

Discipline groups are increasingly recognized as significant stakeholders in digital curation. There is a growing understanding that curation requires significant domain knowledge and that curators with this domain knowledge are more effective than those who do not possess it because they understand and can take account of the diversity of disciplinary cultures and approaches to research. There is, however, a major shortage of domain experts with digital curation expertise, as noted in Chapter 2.

Data Creators

Scientists, scholars, and researchers, singly or as members of research teams, are all involved in some of the processes of digital curation. Data creators, if they are attuned to curation requirements, ensure that the data they bring into being are structured and documented to maintain their longevity and reusability. For digital objects to be usable and reusable, they must be of high quality, well-structured, and adequately documented (Chapter 10 notes these characteristics in more detail). The best time to ensure that these characteristics are present is when the data are created. Their presence makes use and reuse of the data significantly easier. If they are missing, use and reuse becomes exponentially more difficult, if not impossible.

These criteria do not apply only to science data. All digital objects, no matter in what context, will stand a better chance of being long-lived and usable in the future if attention is paid, at the time of their creation, to their structure and to the metadata associated with them.

Adequate documentation is particularly critical for digital curation. To take an example from experimental science—in addition to data sets, information about experiments (such as details of the instrumentation) is usually noted but has not traditionally been considered important for curating the data sets. It is often informally noted in blogs or wikis and is at great risk of being lost (Conway, 2009: 20–21).

Data Users and Reusers

Scientists, scholars, and researchers—in fact anyone who uses and reuses data—are also involved in some of the processes of data curation. Data users and reusers, if aware of digital curation, ensure that any annotations they produce are captured and documented to the extent that they can be understood by other users of those data. Annotation tools are noted in more detail in Chapter 15.

Data Curators

Data curators, whose primary role is managing or looking after data, carry out a wide range of tasks. The tasks of a data curator in the biosciences

context, for example, include ongoing data management, intensive data description, ensuring data quality, collaborative information infrastructure work, and metadata standards work. A fuller list of tasks and responsibilities encompassed by digital curation includes the following:

- Developing and implementing policies and services
- Analyzing digital content to determine what it can be used for
- Providing advice to creators and users/reusers of digital objects
- Ensuring submission of digital objects to a repository
- Negotiating agreements
- Ensuring data quality
- Ensuring that digital objects are structured in the best way to provide access, rendering, storage, and maintenance
- Enabling the use and reuse of digital objects
- Enabling discovery and retrieval of digital objects
- 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 digital objects are viable, able to be rendered, understandable, and authentic
- Promoting interoperability

The DCC website contains transcripts of interviews (Digital Curation Centre, 2004) that provide a helpful overview of the concerns and tasks of data creators, reusers, and curators. A more recent account of the work of digital curators is available in Whyte's (2008) study *Curating Brain Images in a Psychiatric Research Group*. Karasti, Baker, and Halkola (2006) discuss in detail what digital curation involves in the domain of ecological research.

The major shortage of domain experts with digital curation expertise is noted in Chapter 2. One domain, biocuration, is well advanced in both the development of its curation practice and in establishing groups for its professionals. Biocuration is "the activity of organizing, representing, and making biological information accessible to both humans and computers" (Howe et al., 2008: 47–50). In this domain there is a heavy reliance on large, complex databases that are annotated with new data by biocurators, for example, UniPROT (www.uniprot.org) for protein structures. These annotated databases link the data sets with writings on research based on these data and in some domains have become essential to scientists. The curators are typically qualified, experienced scientists who carry out tasks such as extracting knowledge from scientific papers, linking information, inspecting and correcting gene structures and protein sequences, developing and managing controlled vocabularies, ensuring clean data, assisting data users with their research, advising on the design of web-based resources, and encouraging submission of data to databases (Howe et al., 2008: 48). In 2009 a new professional body,

the International Society for Biocuration, was established. Their website (www.biocurator.org) gives further information about this newly formed organization.

CURATE AND PRESERVE: REVIEW

Key points: *Curate and Preserve* is the ongoing process of applying management and administrative actions that curate data.

Key activities:

- Understand why curation is necessary
- Be aware of the range of management and administrative actions applied to curate data
- Apply management and administrative actions to curate data

Summary: Managing Curation

The importance of managing curation throughout the lifecycle of data is acknowledged in the Full Lifecycle Action *Curate and Preserve*. Digital curation aims to produce and manage digital objects with the characteristics of longevity, integrity, and accessibility. The actions required to achieve this aim are the concern of a wide range of stakeholders, from individual data creators to the bodies that fund research.

The next chapter is about another Full Lifecycle Action, *Description and Representation Information*. It examines the importance of metadata (administrative, descriptive, technical, structural, and preservation metadata), standards for metadata, and why metadata is essential for adequate description and control of data over the long term. It also describes representation information, which is required to understand and render both the digital object and its associated metadata.

References

Conway, Esther. 2009. *Curating Atmospheric Data for Long Term Use: Infrastructure and Preservation Issues for the Atmospheric Sciences Community.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/webfm_send/112 (accessed April 26, 2010).

Digital Curation Centre. "Case Studies." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/resources/case-studies (accessed April 26, 2010).

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

———. 2004. "Interviews." Edinburgh: Digital Curation Centre (September–October 2004). Available: www.dcc.ac.uk/community/interviews-setting-scene (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 (November 2008). Available: www.dpconline.org/advice/digital-preservation-handbook.html (accessed April 26, 2010).

Howe, Doug, Maria Costanzo, Petra Fey, Takashi Gojobori, Linda Hannick, Winston Hide, David P. Hill, Renate Kania, Mary Schaeffer, Susan St Pierre, Simon Twigger, Owen White, and Seung Yon Rhee. 2008. "The Future of Biocuration." *Nature* 455 (4 September): 47–50.

Karasti, Helena, Karen S. Baker, and Eija Halkola. 2006. "Enriching the Notion of Data Curation in E-science: Data Managing and Information Infrastructuring in the Long Term Ecological Research (LTER) Network." *Computer Supported Cooperative Work* 15: 321–358.

McGarva, Guy. 2006. "Curating Geospatial Data." Edinburgh: Digital Curation Centre (April 5, 2006). Available: www.dcc.ac.uk/resources/briefing-papers/introduction-curation/curating-geospatial-data (accessed April 26, 2010).

National Library of Australia. "Recommended Practices for Digital Preservation." Canberra, ACT: National Library of Australia. Available: www.nla.gov.au/preserve/digipres/digirespractices.html (accessed April 26, 2010).

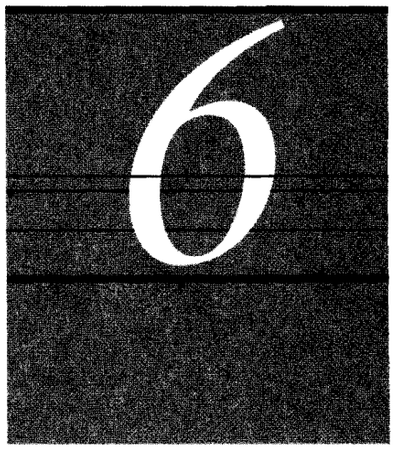
Pennock, Maureen. 2006. "Curating E-science Data." Edinburgh: Digital Curation Centre (August 25, 2006). Available: www.dcc.ac.uk/resources/briefing-papers/introduction-curation/curating-e-science-data (accessed April 26, 2010).

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).

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).

Whyte, Angus. 2008. *Curating Brain Images in a Psychiatric Research Group: Infrastructure and Preservation Issues.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/webfm_send/111 (accessed April 26, 2010).

Description and Representation Information



This chapter notes *Description and Representation Information*, one of the four Digital Curation Centre (DCC) Curation Lifecycle's Full Lifecycle Actions. The activities that comprise this action are stated in the DCC Curation Lifecycle Model in these words: "Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata" (Digital Curation Centre, 2008a). What does this mean in practice?

Higgins (2006) notes, "Metadata is the backbone of digital curation. Without it a digital resource may be irretrievable, unidentifiable or unusable." Effective digital curation relies heavily on information that is added to or associated with the digital object that is being preserved. These additions and associations occur at all points in the curation lifecycle. Maintaining just the data or the digital objects will not ensure that they are preserved, nor will it mean that they will be usable in the long term. What is needed is additional information about the data or digital objects. The key to identifying, locating, accessing, understanding, and using data and digital objects in the future is to add information about them that provides a context and, if required, provides the tools to use those data. This added information comes in several varieties:

- Metadata that describes digital objects and refers to where they are located
- Metadata that provides the technical information needed to use the digital objects
- Metadata that describes what has happened to the digital objects as they move through the curation lifecycle

The DCC Curation Lifecycle Model refers to this added information as *description information* and *representation information*. Description information is added "to ensure adequate description and control over the long term"; representation information is added "to understand and render both the digital material and the associated metadata" (Digital Curation

IN THIS CHAPTER:

Centre, 2008a). The terms *description information* and *representation information* come from the Open Archival Information System (OAIS) Reference Model (see Chapter 3). Essential to OAIS is the concept of information package—a container of the digital object plus information about that object (i.e., description and representation information), which, together, allow the object to be understood and used. The OAIS Reference Model defines the kinds of description and representation information that is needed to archive and use digital objects over the long term.

The Curation Lifecycle Model's statement about the Full Lifecycle Action *Description and Representation Information* indicates that for descriptive information (or metadata) there are different kinds: administrative, descriptive, technical, structural, and preservation metadata. These types of metadata need to be constructed according to defined and accepted standards, with the aim of making sure that there are adequate descriptions of the data or digital objects to enable them to be located and controlled during the period they are curated. Representation information is different from descriptive information (the metadata) that we add to or associate with data or digital objects. Rather than allowing us to manage the data or digital objects, the role of representation information is to help us to understand data or digital objects and to render them (i.e., display, play, or otherwise use them in ways originally intended).

A further distinction between the two categories of added information—description information and representation information—is significant. The added information is developed in different ways, depending on what kind it is. It may be assigned by people (such as catalogers, curators, or data creators), or it may be automatically generated. For example, descriptive metadata is typically assigned, whereas technical metadata is typically derived from data using a computer program. This distinction is significant because constructing description and representation information is resource intensive when carried out manually, so automating this process wherever possible is considered essential in handling large quantities of data.

The key activities associated with the Description and Representation Information action are:

- appreciating the need for description and representation information;
- being aware of where description and representation information is required;
- understanding the key standards that exist for description and representation information; and
- developing policies for applying description and representation information.

The Need for Description and Representation Information

Curating digital objects is not about just the data that comprise the objects. It requires that information about the digital objects is generated

and is also curated. Description and representation information allows the curator to:

- persistently identify digital objects;
- maintain reliable links to the digital objects;
- clearly describe what the digital objects are;
- clearly identify the technical characteristics of the data comprising the digital objects;
- identify who is responsible for the management and preservation of the digital objects;
- describe what can be done to the digital objects;
- describe what is needed to re-present the digital objects at the standard required by users;
- record the history of the digital objects; and
- document the authenticity of the digital objects.

This information also allows users to understand the context of digital objects and their relationship to other digital objects (National Library of Australia, accessed 2010).

The converse situation—when there is insufficient and appropriate description and representation information—is that digital objects can neither be effectively managed nor accessed by those who seek to use them in the future. We may be unable to locate digital objects because the metadata is not specific enough to differentiate among them or because the location of a digital object has changed and there is no record of where it was moved to. We may not be able to read and use the digital objects because we lack sufficient knowledge about their technical properties or structure. We may not be able to understand the digital objects because we don't know enough about the context in which they were created. We may not be able to verify that the data comprising the digital objects are authentic and that we can rely on them (Cunningham, 2008).

Definitions

Administrative, descriptive, technical, structural, and preservation metadata have already been briefly mentioned. It is useful to define these and other terms associated with the Description and Representation Information action.

Description information is descriptive metadata that describes a digital object. *Metadata* is "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (National Information Standards Organization, 2004: 1). For our purposes, it comprises the following:

- **Administrative metadata**: metadata related to the use, management, and encoding processes of digital objects over a period of time. Includes the subsets of technical metadata, rights management metadata, and preservation metadata.

- **Descriptive metadata**: metadata that describes a work for purposes of discovery and identification, such as creator, title, and subject.
- **Technical metadata**: a form of administrative metadata dealing with the creation or storage encoding processes or formats of the resource.
- **Structural metadata**: metadata that indicates how compound objects are structured, provided to support use of the objects.
- **Preservation metadata**: administrative metadata dealing with the provenance of a resource and its archival management. (National Information Standards Organization, 2004: 15–16)

This list can be expanded with respect to curation activities. Using metadata we can note information about the controlled vocabularies applied to a digital object to classify or index its content, about other digital objects it is related to, about what processes (and hardware and software) were used to produce it, and about what processes are required to use it. There is much more: metadata records information we need for curation purposes about intellectual property rights associated with digital objects, about how they were acquired and where from, about users (who can access them, who has accessed them), about which version of the digital object is to hand, about checks of data integrity (such as checksum calculations), and about the preservation actions applied to the digital object, such as migrations. Metadata is also used to record information about the metadata itself, such as how it was created or when it was altered (Higgins, 2006).

Representation information is different. It is "either information which describes how to interpret a data object (such as a format specification), or a component of a technical environment which supports interpretation of that object (such as a software tool or hardware platform)" (Adrian Brown, cited in Rusbridge, 2008). Representation information is explained in more detail later in this chapter.

Standards for Description and Representation Information

Standards are essential for description and representation information. As indicated in other chapters (and especially in Chapter 8), data sharing and reuse, which are at the heart of digital curation, requires interoperability, which in turn requires adherence to standards. Standards that apply to description and representation information are no exception to this requirement.

Standards for description and representation information are listed and described on several websites. The DCC DIFFUSE Standards Frameworks (www.dcc.ac.uk/diffuse) include several subsections, such as standards for metadata content, metadata description, metadata structure, and thesauri and word lists. The large number of standards relevant to this topic is indicated by the further subdivision into standards for authentication, authorities, metadata content, metadata description, metadata structure, reference models and frameworks, searching protocols,

thesauri and word lists, and XML DTD and schemas. The "Data Documentation & Standards" topic on the Preserving Access to Digital Information (PADI) website (www.nla.gov.au/padi/topics/29.html) notes many standards relevant to description and representation information. The Library of Congress has a "Standards" page (www.loc.gov/standards) that is also helpful.

Description Information

Metadata (which can be equated with description information) is defined formally as "structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource" (National Information Standards Organization, 2004: 1). Digital objects without associated description information are impossible to curate, because not enough is known about them to ascertain what they are or whether it is worth expending resources on their curation.

Description information for curation purposes comes in several varieties:

- Metadata that describes digital objects and their location (the *descriptive metadata* and *structural metadata* noted in the Curation Lifecycle Model)
- Metadata that provides the technical information needed to use digital objects (the *technical metadata* noted in the Model)
- Metadata that describes what has happened to digital objects as they move through the curation lifecycle (the *administrative metadata* and *preservation metadata* in the Model).

Metadata that describes digital objects and indicates where they are located consists of descriptive and structural metadata. *Descriptive metadata* is information that allows digital objects to be identified so they can be linked with requests. The name of the creator of the data set and the author of a document are examples of this type of information. This category also includes *structural metadata*, which describes how compound digital objects are organized.

Metadata that provides the technical information needed to use digital objects is *technical metadata*: information about technical characteristics of them such as their format, compression or encoding algorithms, encryption and decryption keys, or software (including the release number) used to create or update the data. Technical metadata also includes information about the overall system environment: the hardware, operating systems, and application software in which the data were created. This metadata enables digital objects to be identified and processed.

Metadata that describes what has happened to digital objects as they move through the curation lifecycle consists of administrative and preservation metadata. *Administrative metadata* is information about the use, management, and encoding processes of digital objects over a period of time; in other words, lifecycle data. It includes information about the creation of digital objects, subsequent updates, transformation,

versioning, summarization, and descriptions of migration and replication. This kind of metadata is necessary to enable data, databases, and digital objects to be managed effectively.

Definitions of the different varieties of metadata vary according to the context and there is sometimes overlap, especially about what kinds of metadata are described as preservation metadata. This is illustrated in Figure 6.1.

Preservation Metadata

Preservation metadata is defined on the PADI website as "structured ways to describe and record information needed to manage the preservation of digital resources." It is used to store "technical details on the format, structure and use of the digital content, the history of all actions performed on the resource including changes and decisions, the authenticity information such as technical features or custody history, and the responsibilities and rights information applicable to preservation actions" (National Library of Australia, 2007).

Preservation metadata is essential for ensuring the long-term accessibility of digital objects. It does this by providing a mechanism to record information about the digital objects so that they are described in sufficient

Figure 6.1. Description Information and Its Functions

	Broad Function	Type	Specific Function	Examples
Descriptive Information	Describes data and their location	Descriptive metadata	Allows data to be identified so they can be linked with requests	Name of the creator of the data set Name of the author of a document
		Structural metadata	Describes how compound digital objects are organized	Relationship of TIFF page image to other page images
	Provides the technical information needed to use data	Technical metadata	Provides the technical information needed to use data	Format Compression or encoding algorithms Encryption and decryption keys Software (including release number) used to create or update the data
			Provides information about the overall system environment	Hardware, operating systems, application software in which the data were created
	Describes what has happened to data as they move through the curation lifecycle	Administrative metadata	Provides information about the use, management, and encoding processes of digital objects over a period of time	Information about data creation, subsequent updates, transformation, versioning, summarization Descriptions of migration and replication
		Preservation metadata	Records the preservation actions that have been applied to data over time	File format Significant properties Technical environment Fixity information

detail to identify them and to record information about the requirements that must be met to use them. It also provides a way of recording the preservation actions that have been applied to the digital objects over time. More specifically, preservation metadata:

- Identifies the material for which a preservation programme has responsibility
- Communicates what is needed to maintain and protect data
- Communicates what is needed to re-present the intended object (or its defined essential elements) to a user when needed, regardless of changes in storage and access technologies
- Records the history and the effects of what happens to the object
- Documents the identity and integrity of the object as a basis for authenticity
- Allows a user and the preservation programme to understand the context of the object in storage and in use. (UNESCO, 2003: Section 14.20)

What kind of information is commonly recorded in preservation metadata? The file format of a digital file is one kind. This often consists of a reference to a unique identifier for the format in a format registry (format registries are noted in Chapter 13). Significant properties (the properties or characteristics of a digital object that must be maintained over time; see Chapter 10) are another. An important kind of preservation metadata is environment information, recording the technical environment in which a digital object is used, such as the hardware, software, and other files required to use it. Fixity information, indicating whether a file has changed, is recorded usually in the form of a checksum calculation and the algorithm used to produce it. (This and the following paragraph are based on Priscilla Caplan's [2006] installment on preservation metadata in the *DCC Curation Reference Manual*, recommended for further study of preservation metadata.)

Technical metadata, another kind of preservation metadata, describes the technical characteristics of files and bit streams. Some of these characteristics are required for all files, for example, size, format, and fixity information. Other characteristics are relevant only to specific file types or formats; audio files, for example, require information about duration of the sound recording, sampling frequency, bit rate, compression, and the number of tracks and their relationships. Details of a digital object's provenance (see Chapter 15) are also recorded in the preservation metadata. Provenance information could include where a digital object originated and who has had custody of it, who created it, who owns the rights to it, and what preservation actions have been applied to it (with details of what the preservation action was, when it was applied, who and what was involved, and the result). The way in which metadata is associated with the content (digital object) to create an information package is also recorded in preservation metadata.

All digital curation actions require metadata. Providing adequate quantities of appropriate metadata is a major challenge. The ideal is to collect metadata by automated processes close to the point of data creation

A **format registry** is an online listing of definitive technical information about file formats. Format registries play an important role in supporting long-term access to digital objects. Examples of format registries include the Global Digital Format registry (www.gdfr.info) and PRONOM (www.nationalarchives.gov.uk/PRONOM), in 2010 in the process of being combined to form the Unified Digital Format Registry (www.udfr.org).

Automated software tools for extracting technical metadata from digital objects, and tools for converting this extracted metadata into XML schema elements, are available. Examples include the National Library of New Zealand's Metadata Extraction Tool, and the Ecological Metadata Language editor. These are noted further in Chapter 13.

so that the need for costly human input is minimized. For example, all the information about an experiment, the environment, the people, as well as the data flowing from the equipment and samples should ideally be captured through automated processes without human intervention. Automated software tools for extracting technical metadata from digital objects, and tools for converting this extracted metadata into XML schema elements, are available and are noted in Chapter 13.

Persistent Identifiers

Persistent identifiers are labels for digital objects that remain the same regardless of where the object is located. They allow us to refer to a digital object and locate it even when it moves to other servers or to other repositories or archives. Reliable identification of data and digital objects is essential for providing long-term access to them and ensuring their reliability and authenticity, thus enabling their reuse over time. We need to know where to find the data or digital object, and persistent identifiers help us do so. If a digital object is moved, it can always be located through its persistent identifier, which does not change. A persistent identifier is also a unique identifier, one that refers to only a single digital object and can, therefore, help reduce confusion if there are several versions of a resource.

Much of the discussion of persistent identifiers has been about their application to web resources. The normal way of identifying web material is to use a Uniform Resource Locator (URL). However, a URL points to only one web location and, if the location of material at the URL changes (e.g., if the material is moved from one domain name to another), the URL changes and the material becomes inaccessible. Many schemes exist for application to web resources. Those encountered most frequently include the Uniform Resource Name (URN), the Persistent Uniform Resource Locator (PURL), and the widely used Digital Object Identifier (DOI) which has achieved National Information Standards Organization (NISO) standard status as ANSI/NISO Z39.84. The "Persistent Identifiers" page on the PADI website provides a summary of the major persistent identifier schemes for web resources and gives a comprehensive list of publications about the topic (National Library of Australia, 2002). The use of persistent identifiers is not limited to web material; it is equally essential for linking and citation of primary research with data sets. (Two good sources of further information about persistent identifiers have been written by Joy Davidson [2006] and Emma Tonkin [2008].)

Metadata Schemas and Standards

Metadata schemas are the formal expressions of the individual kinds of metadata (or elements). These are often adopted as metadata standards that, because they provide standardized ways of selecting metadata and

of expressing it, allow the results to be shared among more than one organization. Sharing is essential for effective curation, as noted elsewhere in this book (especially in Chapter 8). Using standards for metadata ensures that the description information created is consistent and indicates the elements that are required to manage and curate the digital objects they describe.

Metadata standards provide standards in three areas: structure, semantics, and syntax. How the metadata is structured is indicated by noting the elements that are described; the semantics describe the meanings of each element; and the syntax indicates how to write the metadata.

Different kinds of metadata standards are available and are often used in combination in curation contexts. Some of the widely applied standards for preservation metadata are noted next. There are many such standards, some of which are discipline specific. The aim here is to indicate examples to illustrate how they work, not to offer a comprehensive listing.

One characteristic of many of the metadata standards used in digital curation is that they are XML based. Using standards based on XML has numerous advantages. XML is an open (nonproprietary), well-supported, and widely adopted standard for encoding textual data, designed to be used regardless of the hardware platform. It is well supported by open software applications, and, because it has been in use since 1996 and was based on SGML (Standardized General Markup Language), an ISO standard that predates XML by ten years, it is widely understood. The use of XML allows metadata to be exchanged and reused by different archives because the way the metadata is expressed is standardized. Other advantages are that data in XML are human readable and structured hierarchically, making them easier to check if the data become corrupted. Reese and Banerjee (2008: 97–107) provide an informative section in their book on this topic titled "Why Use XML-Based Metadata?"

The most common metadata standards applied to digital curation are Preservation Metadata: Implementation Strategies (PREMIS), Metadata Encoding and Transmission Standard (METS), Metadata Object Description Schema (MODS), and Metadata Authority Description Schema (MADS), noted in more detail in the following sections. These are all related to XML, although in different ways. PREMIS exists independently of XML but is often used in conjunction with METS, which is a standard for expressing descriptive information in XML. MODS and MADS combine XML and guidelines for developing metadata. Underlying all of these standards is XML, which allows the metadata developed to be expressed in a standardized way.

PREMIS

Several preservation metadata schemas have been developed, PREMIS being one of the most widely adopted. The *PREMIS Data Dictionary for Preservation Metadata* (2008) defines a core set of preservation metadata elements that has wide applicability in the preservation community. It defines preservation metadata as metadata that:

- supports the viability, renderability, understandability, authenticity, and identity of digital objects in a preservation context;
- represents the information most preservation repositories need to know to preserve digital materials over the long-term;
- emphasizes "implementable metadata": rigorously defined, supported by guidelines for creation, management, and use, and oriented toward automated workflows; and
- embodies technical neutrality: no assumptions made about preservation technologies, strategies, metadata storage and management, etc. (*PREMIS Data Dictionary for Preservation Metadata*, 2008: 1)

The PREMIS data model defines relationships between digital preservation "entities":

- **Intellectual Entities**: not defined in PREMIS. Users are expected to apply other relevant metadata standards. Examples are a particular book, map, photograph, database, or website.
- **Objects** (divided into three types: representation, file, and bit stream): what the repository preserves. Examples are a PDF file, an audio stream in uncompressed PCM, and a video stream in MJPEG.
- **Events**: actions on an object in the preservation repository. These document provenance and track the history of the object. Examples are the action of verifying that a file is well formed and creating a new version of an object in a more contemporary format (i.e., migration).
- **Agents**: people, organizations, or software programs associated with preservation events in the life of an object.
- **Rights**: an agreement with a rights-holder that allows a repository to take actions in relation to objects in the repository. An example is an organization giving permission to make copies of an object.

PREMIS is used in conjunction with other applicable metadata standards where appropriate. The PREMIS schema has been endorsed for use with METS. PREMIS maintenance is sponsored by the Library of Congress.

METS

METS (www.loc.gov/standards/mets) is a standard for encoding in XML the metadata describing or characterizing digital objects. It provides a means of associating all the metadata about a digital object with the object—that is, it is a "container format" specifying how different kinds of metadata can be packaged together (Caplan, 2008: 16). METS was developed with reference to the OAIS Reference Model's information package concept. The Library of Congress supports METS by acting as its maintenance agency.

METS encourages interoperability by providing a standard for exchanging digital materials among institutions. As the quantity of digital

objects being curated increases, so too does the amount of metadata required for curation. When digital objects are shared between repositories, the metadata about them is also shared, so a common data transfer standard greatly increases the effectiveness of this sharing process. METS is designed to provide this interoperability.

A METS XML document has five major sections (based on the Library of Congress's "METS: An Overview & Tutorial," accessed 2010):

- **Descriptive metadata**: contains pointers to external descriptive metadata (e.g., a MAchine Readable Cataloging [MARC] record in a library catalog) or contains descriptive metadata, or both.
- **Administrative metadata**: provides information about how the files were created and stored, intellectual property rights, the original source object, and the provenance of the files comprising the digital object.
- **File groups**: lists all files comprising all versions of the digital object.
- **Structural map**: outlines a hierarchical structure of the digital object, linking the parts of that structure to content files and metadata about each element.
- **Behavior**: used to associate executable behaviors with content in the METS object.

MODS and MADS

MODS (www.loc.gov/standards/mods) is a standard developed by the Library of Congress for encoding descriptions of information resources. It consists of a bibliographic element set expressed in XML so that metadata developed using this schema can be readily shared. MODS is designed to allow the importing of existing catalog data that is in MARC 21 format, so it is especially valuable if there is legacy metadata in MARC format that needs to be moved to another metadata standard.

Associated with MODS is MADS (www.loc.gov/standards/mads). MADS is also an XML schema, developed to complement MODS and to use data in MARC 21 format. It is used to describe and provide authority control for names of people, organizations, events, and terms. MADS can be used independently of MODS but is designed to work well with it.

Metadata standards are often used in combination in curation contexts. Dappert and Enders (2008) illustrate how three metadata standards, METS, MODS, and PREMIS, each with different aims, work together in a system for archiving e-journals. Some elements can be expressed in more than one standard; for example, both METS and PREMIS provide for file format information, but PREMIS allows linking to the PRONOM format registry while METS does not. Both were used. Dappert and Enders (2008) conclude that "no single existing metadata schema accommodates the representation of descriptive, preservation and structural metadata."

However, the use of XML-based standards such as METS, PREM and MODS cannot be assumed in all contexts. There are many meta standards in widespread use that have been developed for specific c munities. As one example, geospatial data uses metadata schemas develo specifically for it, such as ISO 19115 *Geographic Information—Me data* and extensions based on it, but these do not accommodate sc preservation metadata. In particular, these geospatial metadata standa "do not provide a wrapper function that would allow additional techn or administrative metadata elements to be associated with . . . the d producer-originated metadata," such as administrative metadata not information about the acquisition of the data, technical metadata rela to data integrity, or metadata recording preservation actions applied the data by the archive. The geospatial community has not to d sought to combine metadata developed according to its standards w preservation metadata based on the METS and PREMIS standa (McGarva, Morris, and Janée, 2009: 22–25).

Representation Information

Representation information is a key concept in curation but one that "often misunderstood" according to Rusbridge (2007). Representati information is the information required to make a bit stream retrieval as a meaningful digital object. It describes how to interpret a digi object (such as a format specification) or a component of a techni environment that supports interpretation of that object (such as a softw tool or hardware platform; as noted by Chris Rusbridge [2008] i Digital Curation Blog posting).

Representation information is added to data and digital objects so can understand and render both the digital material and the associat metadata. We need more than just the information about the file form to ensure that the bit stream is accessible and understandable over ti We also need information about operating system and hardware depe encies, character encoding, and algorithms, among other things. Witho representation information, digital objects are just a sequence of me ingless binary information. As the DCC puts it:

> Digital objects are stored as bitstreams which are not understandable to a human being without further data to interpret them. Representation information is the extra structural or semantic information which converts raw data into something more meaningful. For example, structure information could tell a computer to interpret a string of bits as ASCII characters, and semantic information could explain what a particular mathematical symbol means. (Digital Curation Centre, accessed 2008b: Q6)

A case study of curation in the eCrystals Data Repository emphasi the need for adequate representation information. Representation inf mation is collected and maintained because it reduces the risks of being able to understand information in the future. Documentation data formats, software, standards, and programming languages t

become obsolete is needed, as is documentation of the specialized knowledge about how to manipulate the data (Patel and Coles, 2007). Another example of representation information (RI) is provided by Rusbridge (2007), who asks us to "imagine a social science survey dataset encoded with SPSS [common statistical software]. We may have all the capabilities required to interpret SPSS files, but still not be able to make sense of the dataset if we do not know the meaning of the variables, or do not have access to the original questionnaires. Both the latter would qualify as RI."

OAIS and Representation Information

The concept of representation information is derived from the OAIS Reference Model (see Chapter 3), which categorizes the information required for preservation as Content Information, Representation Information, Preservation Description Information (broken down into Reference, Context, Provenance, and Fixity Information), and Packaging Information. The representation information category is divided into three classes: Structure Information, Semantic Information, and Other Representation Information.

Representation information is required to "understand and render" the digital material. How is "understandable" defined? Understandable to whom? The answer to these questions lies in the OAIS Reference Model. The information being preserved should be "independently understandable to the Designated Community"; that is, members of the Designated Community do not require expert assistance to understand and use the information of that community. The archive needs to understand the knowledge and requirements of the Designated Community in order to ensure that appropriate representation information is available. This is further examined in "Appendix 2: Understandability & Use" in *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (RLG-NARA Task Force on Digital Repository Certification, 2007: 77–80).

Representation information takes various forms. *Structure Information* describes the formats and data structures relevant for processing and rendering digital objects. Structure information is usually information about file formats. File formats used for processing are often open standards, and formal descriptions of these formats assist in automated processing. File formats for rendering are, by comparison, more likely to be commercially based and proprietary formats, making it more difficult or impossible to record their structure. If the file formats and data structures of digital objects created by proprietary software cannot be recorded, it may be necessary to keep the original software or other software that provides access to the digital object.

Semantic Information provides additional information about the content of a digital object, particularly information that defines relationships among objects or parts of objects. Examples include data dictionaries, ontologies, and thesauri. For a spreadsheet, for example, semantic information could indicate which units of measurement have been used for which data values.

Other Representation Information is any other kind of representatic information thought necessary to interpret the digital object. It coul be information about relevant software, hardware, and storage medi encryption or compression algorithms, or documentation. Time dependent information is another example; some data sets change ov time, and it may be important to record the state of a data set at specif points in time, such as on a particular date or when a specific action h occurred.

Two examples illustrate these points. The first is of a binary fil produced by Portable Document Format (PDF) software, containin information in English describing a medical procedure. For this th Designated Community is defined as having a knowledge base typical c second-year medical students who read English; if they are to understan the file they need to see it rendered in the same way it was rendere when it was originally submitted to the archive. To achieve this, the rep resentation information needed is information about the PDF-A format The archive curates and provides to the Designated Community th binary file, preservation description information about the file, and PDF-A rendering application. The second example is of a musical scor for a synthesizer, in a nonproprietary binary format. The Designate Community is defined as German readers who wish to generate musi from a digital representation of the score, using a computer and synthe sizer. To understand the score they need the binary bit stream, plu information about the PDF-A format, both in German. In both case the archive may not itself curate the PDF-A specifications and renderin application but may instead provide a link to a repository where thes are available. (These examples come from "Appendix 2: Understandabilit & Use" in *Trustworthy Repositories Audit & Certification: Criteria an Checklist* [RLG-NARA Task Force on Digital Repository Certificatio 2007: 77–80] where more examples can be found.)

Sharing Representation Information

Representation information needs to be curated in its own right. needs to be stored in association with the digital object and other met data so that the digital object can be understood and rendered in th future. It may be stored by the repository that stores the digital objec or stored externally by another reliable repository. Much representatio information is not unique and can be applied to many digital objects– it can be reused. One way to do this is to establish repositories of repr sentation information. These help to reduce duplication of effort i developing representation information and to share knowledge an expertise (Brown, 2008: 10).

Collecting and maintaining adequate representation information is major endeavor. One example of a repository to assist this undertaking the Representation Information Repository established by the Cultura Artistic and Scientific Knowledge for Preservation, Access and Retriev (CASPAR) project and the DCC (registry.dcc.ac.uk:8080/Registr Web/Registry/index.jsp). Representation information can be deposite

in the registry so that it becomes an authoritative source of representation information for the data curation community. Format registries such as PRONOM (developed by The National Archives [United Kingdom]) and the Unified Digital Format Registry (formed by the merger of the Global Digital Format Registry and PRONOM from 2010) can also be considered as representation information repositories. They provide details about file formats, which are essentially the Structure Information category of representation information. Adrian Brown's (2008) paper is a useful source of further information.

Policies for Description and Representation Information

As is the case for all aspects of curation, policies are needed to guide the use of description and representation information. The key aspects that policy should address are the determination of who has access to description and representation information and under what conditions, reuse, the kinds of description and representation information that are required, and the metadata standards and schemas that will be applied. (This section is based on *Policy-making for Research Data in Repositories: A Guide* [Green, Macdonald, and Rice, 2009], a compilation of considerations that have been aggregated from numerous sources, as stated and listed in the guide.) A policy relating to who has access to description and representation information and under what conditions might, for example, specify that description and representation information is available without charge to anyone. A policy about reusing description and representation information could consider whether prior permission is needed, whether it can be reused for commercial purposes, or whether it should be made available for harvesting by other organizations.

Policies and their associated procedures for the kinds of description and representation information that are required and where such information is drawn from are likely to be very detailed. They could include statements about what description and representation information must be supplied to the archive and by whom; for example, the data creator may be required to provide a codebook for statistical data, or a format specification, or an explanation of the research protocol or methodology. Procedures that specify what metadata standards are used in the provision of description information might, for example, prescribe that descriptive metadata is created using MODS.

Summary: Curation Needs Metadata

Description and representation information is commonly considered to be essential for curation but is also, perhaps, the factor that restricts effectiveness the most. There are limits to the human resources that can be applied to curation, given the massive quantities of data that are involved, so automation of as many of the curation processes as possible

DESCRIPTION AND REPRESENTATION INFORMATION: REVIEW

Key points: *Description and Representation Information* is the ongoing process of assigning description and representation information at all stages of the curation lifecycle.

Key activities:

- Appreciate the need for description and representation information
- Be aware of where description and representation information is required
- Understand the key standards that exist for description and representation information
- Develop policies for applying description and representation information

is necessary. Metadata is critical to support this high level of automation. Ideally it is automatically created when the data are created and is then available when the data are ingested into the long-term archive. There has, however, been very little progress toward achieving this ideal.

Much effort has been expended on aspects of description and representation information, but there is more to be done. Caplan's (2006: 23) comments about preservation metadata are apposite:

> Many specifications for preservation metadata have been published and significant progress has been made towards standardizing a core set of preservation metadata elements. However... the success of preservation metadata in supporting long-term preservation is largely untried. The metadata recorded today is our best guess of what will be useful tomorrow. As more experience is gained with various preservation strategies and different preservation repository systems, we can expect our understanding of preservation metadata to grow increasingly more sophisticated in the future.

Librarians and archivists are very familiar with the need for metadata to support curation activities, but for curating digital objects the requirements are more extensive, especially those for preservation metadata. For individuals who create digital objects, metadata is equally important. Consider, for example, the difficulty of searching through hundreds of digital photographs labeled JPEG1, JPEG2, JPEG3, and so on.

The next chapter describes *Preservation Planning*, another Full Lifecycle Action. It notes the need for planning for preservation throughout the curation lifecycle of data and digital objects and also notes the importance of policy for curation.

References

Brown, Adrian. 2008. *White Paper: Representation Information Registries.* London: PLANETS. Available: www.planets-project.eu/docs/reports/Planets_PC3-D7_RepInformationRegistries.pdf (accessed April 26, 2010).

Caplan, Priscilla. 2006. *DCC Digital Curation Manual: Instalment on Preservation Metadata.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/preservation-metadata (accessed April 26, 2010). Used by permission of Priscilla Caplan.

———. 2008. *The Preservation of Digital Materials* (Library Technology Reports 44, no. 2). Chicago: ALA TechSource.

Cunningham, Adrian. 2008. "The Uses of Metadata in Public Administration." Glasgow: Digital Preservation Europe (November 7, 2008). Available: www.digitalpreservationeurope.eu/publications/briefs/uses_of_metadata_in_public_administration.pdf (accessed April 26, 2010).

Dappert, Angela, and Markus Enders. 2008. "Using METS, PREMIS and MODS for Archiving eJournals." *D-Lib Magazine* 14, no. 9/10 (September/October). Available: www.dlib.org/dlib/september08/dappert/09dappert.html (accessed April 26, 2010).

Davidson, Joy. 2006. "Persistent Identifiers." Edinburgh: Digital Curation Centre (October 18, 2006). Available: www.dcc.ac.uk/resources/briefing-papers/introduction-curation/persistent-identifiers (accessed April 26, 2010).

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

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

Green, Ann, Stuart Macdonald, and Robin Rice. 2009. *Policy-making for Research Data in Repositories: A Guide.* Version 1.2. Edinburgh: EDINA and University Data Library. Available: www.disc-uk.org/docs/guide.pdf (accessed April 26, 2010).

Higgins, Sarah. 2006. "What Are Metadata Standards?" Edinburgh: Digital Curation Centre (August 30, 2006). Available: www.dcc.ac.uk/resources/briefing-papers/standards-watch-papers/what-are-metadata-standards (accessed May 6, 2010).

Library of Congress. "METS: An Overview & Tutorial." Washington, DC: Library of Congress. Available: www.loc.gov/standards/mets/METSOverview.v2.html (accessed April 26, 2010).

McGarva, Guy, Steve Morris, and Greg Janée. 2009. *Preserving Geospatial Data* (Technology Watch Report). York: Digital Preservation Coalition. Available: www.dpconline.org/technology-watch-reports/download-document/363-preserving-geospatial-data-by-guy-mcgarva-steve-morris-and-gred-greg-janee.html (accessed April 26, 2010).

National Information Standards Organization. 2004. "Understanding Metadata." Bethesda, MD: NISO Press. Available: www.niso.org/publications/press/UnderstandingMetadata.pdf (accessed April 26, 2010). Reprinted with permission from the National Information Standards Organization (NISO). Original text © NISO Press, 2004.

National Library of Australia. "Digital Preservation: Critical Elements of Preserving Digital Collections." Canberra, ACT: National Library of Australia. Available: www.nla.gov.au/preserve/digipres/elements.html (accessed April 26, 2010).

———. 2002. "PADI—Persistent Identifiers." Canberra, ACT: National Library of Australia. Available: www.nla.gov.au/padi/topics/36.html (accessed April 26, 2010).

———. 2007. "PADI—Preservation Metadata." Canberra, ACT: National Library of Australia (July 2007). Available: www.nla.gov.au/padi/topics/32.html (accessed April 26, 2010). Used by permission of the National Library of Australia.

Patel, Manjula, and Simon Coles. 2007. *A Study of Curation and Preservation Issues in the eCrystals Data Repository and Proposed Federation.* Bath: UKOLN. Available: www.ukoln.ac.uk/projects/ebank-uk/curation/eBank3-WP4-Report (Revised).pdf (accessed April 26, 2010).

PREMIS Data Dictionary for Preservation Metadata. 2008. Version 2.0. Washington, DC: PREMIS Editorial Committee. Available: www.loc.gov/standards/premis/v2/premis-2-0.pdf (accessed April 26, 2010).

Reese, Terry, and Kyle Banerjee. 2008. *Building Digital Libraries: A How-To-Do-It Manual.* New York: Neal-Schuman.

RLG-NARA Task Force on Digital Repository Certification. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Chicago: Center for Research Libraries. Available: www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed April 26, 2010).

Rusbridge, Chris. 2007. "Representation Information: What Is It and Why Is It Important?" Digital Curation Blog, comment posted July 6, 2007. Available:

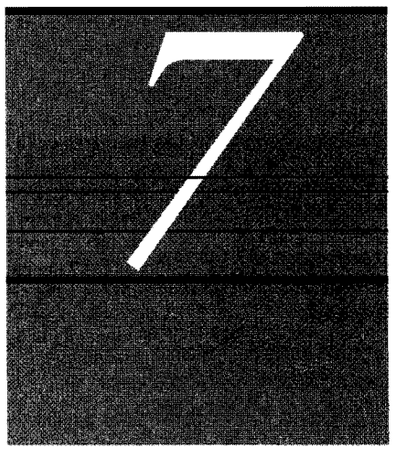
digitalcuration.blogspot.com/2007/07/representation-information-what is-it.html (accessed April 26, 2010).

———. 2008. "Representation Information from the Planets?" Digital Curation Blog, comment posted April 14, 2008. Available: digitalcuration.blogspot.com/2008/04/representation-information-from-planets.html (accessed April 26, 2010). Used by permission of Chris Rusbridge, Digital Curation Centre.

Tonkin, Emma. 2008. "Persistent Identifiers: Considering the Options." *Ariadne* 56 (July). Available: www.ariadne.ac.uk/issue56/tonkin (accessed April 26, 2010).

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). © UNESCO 2003; used by permission of UNESCO.

Preservation Planning and Policy



This chapter investigates *Preservation Planning*, one of the four Full Lifecycle Actions in the Digital Curation Centre (DCC) Curation Lifecycle Model. Preservation planning is the ongoing process of planning data curation activities. The activities contained in the Preservation Planning action are referred to in the DCC Curation Lifecycle Model in these words: "Plan for preservation throughout the curation lifecycle of digital material. This would include plans for management and administration of all curation lifecycle actions" (Digital Curation Centre, 2008). The key activities encompassed by Preservation Planning are:

- appreciating the need for planning at all stages of curation,
- developing plans for all stages of curation, and
- periodically reviewing and updating curation procedures.

This chapter notes the need for planning throughout the curation lifecycle of digital material. It also describes the importance of developing policy for all aspects of digital curation and refers to the findings of recent research into the costs of digital curation.

IN THIS CHAPTER:

Risk Management as the Context for Preservation Planning

To ensure their long-term accessibility, authenticity, and integrity, digital objects need to be managed over the whole of their lifecycle, beginning at the point when they are created. To achieve the active management of data throughout their lifecycle, "constant maintenance and elaborate 'lifesupport' systems' must be planned" (Hedstrom, 2002: 3). Planning is intrinsic to the OAIS Reference Model on which most digital archives are based (see Chapter 3). In the OAIS Reference Model, the *Preservation Planning* function covers the development of preservation strategies, undertaking technology watch and other planning and policy activities.

One way of thinking about the planning process is to consider it as proactive preservation activities aimed at minimizing risks. The active

management of risks over time is essential to preserve digital objects and to ensure that they remain usable in the future. A risk management approach is increasingly common in the preservation of both digital and nondigital materials.

Risk management is aimed at reducing the likelihood that compromising events will occur and at limiting their impact when they do occur. A straightforward example is the reduction of risk from virus attack by using virus protection software and ensuring it is regularly updated. Another example is making regular backups and checking the integrity of the backups to minimize the risks of loss if data become corrupted. Planning for risk reduction activities and their implementation is critical.

One major area where planning is required is during the creation of digital objects—for example, planning to create them in open, well-supported, uncompressed, and stable file formats. (This is noted further in Chapters 9 and 10.) Planning to do this reduces the risk, later in the digital object's life, that file formats will become obsolete, thereby decreasing the odds of being able to use those digital objects in the future.

Broad principles of risk management have become standard practice for digital preservation. They include protecting data by implementing a regular backup procedure; maintaining multiple copies of the bit streams; having disaster recovery contingencies in place; providing secure and stable media storage conditions; copying data to more stable media at defined intervals; and ensuring data security by implementing procedures for virus protection and unauthorized access. To these can be added maintaining ongoing access to digital materials by ensuring sufficient relevant metadata is associated with data sets, possibly limiting the range of formats that the archive manages, community watch activities to monitor for obsolescence and applicable new developments, and collaboration to develop solutions. Applying a risk management approach requires planning at all stages of the curation lifecycle. Planning provides information about digital objects that are at risk, determines the actions required to reduce risk, and ascertains the resources required (Clifton, 2005).

Risk management is not the only context in which planning for digital curation takes place. The requirements of research funding bodies also specify that a data management plan and plans for data sharing, curation and preservation be included in applications for funding. (An indication of these requirements can be seen in "Table 2: Summary of Research Data Policies of Research Funders" of a report on the establishment of a research data service for the United Kingdom [Serco Consulting, 2008: 9–12]). For example, the Wellcome Trust (accessed 2010), a major funder of medical research, expects researchers to "plan at the proposal stage how they will manage and share their data."

RISK MANAGEMENT: MORE EXAMPLES

- Information storage media failure—affects ability to access digital objects
- Equipment obsolescence—affects ability to access digital objects
- Insufficient metadata—makes digital objects more difficult to preserve
- Inadequate ongoing resourcing—threatens the viability of an organization
- Physical disasters—threaten digital objects and storage

Key Planning Steps

Planning for each Sequential Action of the Curation Lifecycle Model, and also for some of the Full Lifecycle Actions, is essential.

Planning for Sequential Actions

The design of data and representation information is planned at the *Conceptualise* stage so that the data can be preserved and reused in optimal ways. Collecting data needs planning (in the *Create or Receive* action of the Lifecycle) to ensure that they, together with relevant description and representation information, are collected in ways that ensure their accuracy. To ensure that decisions are consistent, *Appraise and Select* requires planning and the development of selection and retention policies. Preparing digital objects to add to a digital archive and adding them to a digital archive (the *Ingest* activities) require planning the procedures for assigning persistent identifiers, for virus checking, for authenticity checks such as checksums, and for many other processes. Planning to store digital objects so that they retain their authenticity is essential in the *Preservation Action* stage. In *Store*, planning to develop sustainable models and long-term institutional commitment to preserving digital objects is necessary. Planning to ensure that digital objects are accessible to users and reusers is the basis of the *Access, Use, and Reuse* action; examples are planning to provide multiple access methods to digital objects, to heighten their visibility and enable reuse, and determining metadata that assists in their discovery.

Planning for Full Lifecycle Actions

The *Curate and Preserve* Full Lifecycle Action is concerned with the management and administration of curation and preservation (see Chapter 5) and focuses on actions that ensure the longevity of digital objects, their authenticity, and that they remain accessible. Planning is a prerequisite for all of these. For example, planning is needed to manage multiple copies of a bit stream (where they are stored, how their authenticity is checked, and so on). The *Description and Representation Information* Full Lifecycle Action (see Chapter 6) necessitates planning to ensure that sufficient and appropriate representation information to describe the digital objects and record their storage and manipulation is available, that digital objects are consistently cited according to relevant standards so they can be analyzed and reused, and that relevant metadata are added so the digital objects are discoverable.

Steps to take in planning for curation have been codified. One example is the British Atmospheric Data Centre (BADC) *Data Management Plan Template* (British Atmospheric Data Centre, 2008?). Application of this template helps ensure that the aims of curation are achieved through defining responsibilities, creating a high-quality archive, supporting data creators and users, and adhering to conditions associated with the use and deposition of the data. The template poses questions under these topics:

- Rights and responsibilities
- What is the dataset
- Format of dataset
- Metadata—information about the data

- Ownership of data
- Data archiving
- Storage and backup
- Data distribution
- Access to third-party data
- Publications
- Liaison between data repository and dataset users/program participants

The questions relating to the "Ownership of data" section ask "Who ha ownership of the data? Is it copyrighted/protected? What is the sourc of the data?" In the "Storage and backup" section, the questions are:

> Who is responsible for the integrity of the data? Is this the primary archive of this data or is it mirrored from elsewhere? How difficult would it be to replace and would it be important to replace it? How often should it be backed up? . . . And onto what medium—disc, tape, . . . How long should the archive keep the data? What should we do with it after this time? (British Atmospheric Data Centre, 2008)

Appendixes to the data management plan that is developed by applying the template cover conditions for depositing data in the archive, fo using data from the archive, and for accessing those data; a list of dat sets is also attached to the data management plan. (Chapter 10 is als relevant to planning curation processes.) The DCC also released fo consultation in 2009 a draft template of a "Data Management Pla Content Checklist" (Donnelly and Jones, 2009).

Planning for active management of data and digital objects over th whole of their lifecycle is based on evaluating potential solutions or set of actions against the requirements of an archive, then developing a pla based on this evaluation. To date there are relatively few software tool to assist in this process, so it continues to be largely manual. One softwar tool that has been developed is Plato, an outcome of the Preservatio and Long-Term Access through Networked Services (PLANETS) projec (www.planets-project.eu). The background to the PLANETS project' preservation planning interests and Plato's development has been summa rized by Martin Donnelly (2008). Plato is a decision support tool tha assists in making decisions about which preservation actions best suit th digital objects that planners are interested in preserving. It is available a open-source software (www.ifs.tuwien.ac.at/dp/plato/intro.html).

Policy for Curation

What Policies Address

Developing policies for all aspects of digital curation is vital for its effec tiveness. This section notes the kinds of policies that are required an what they contain. Policies relating to specific actions in the curatio

lifecycle are noted in other chapters: for example, Chapter 9, "Creating Data," includes a section about policies for creating and receiving data.

Policies provide clear, long-term direction and guidance and are regularly reviewed and updated. They provide long-term guidance by unambiguously stating principles, values, and intentions. Their value lies in the clear articulation of these so that expectations are confirmed and explicit and consistent decisions can be made on the basis of the statements they contain.

Having policies about curation in place assists an organization to develop a digital curation strategy and to plan coherent digital curation programs. They ensure and reinforce accountability: for instance, they serve to demonstrate that funds can and will be used responsibly and ensure consistency in this. Other benefits of having policies in place include protecting organizations if they are accused of any wrongdoing, indicating clearly to staff what is acceptable practice and what is not, and stating to the rest of the world that the organization takes its curation responsibilities seriously.

In addition to policies about curation, procedures for curation should be in place. Unlike the general statements in policies, the substance of procedures is very specific. Policies are implemented through procedures, which describe the process of implementing policy and work together with policies to achieve the overall goals of an organization. For example, a policy about ingest might state that digital objects are normalized, that the version that precedes the normalized version is also deposited in the archive, that the object is given a unique identifier and checked for viruses before ingest. The procedures related to this policy will document in more detail what happens: for example, the range of file formats that are accepted, the range of file formats they are normalized to and the steps to take in the normalization process, the steps to take to capture the version before it is normalized, the naming conventions used to provide a unique identifier, and the standards and software used to check for the presence of viruses.

Although the contents of policies will clearly differ according to an organization's mission and requirements, good policies usually have elements in common. They state what is allowed and what is not allowed. They indicate how the policy will be monitored and who has the responsibility for ensuring this. They note links to other relevant policies and to statements about procedures. They also note the date when they are to be reviewed and how frequently review should occur.

The central role of policy can be seen in the *Administration* function of the OAIS Reference Model (see Chapter 3). The *Administration* function is carried out through lower-level functions, one of which is *Establish Standards and Policies.* Based on inputs such as budget information and scope of an organization, this function establishes standards and policies that are implemented by other functions in the OAIS Reference Model. For example, this *Establish Standards and Policies* function develops storage management policies that are implemented by the *Archival Storage* higher-level function and policies about format and documentation standards and procedures that are implemented by the

higher-level function *Ingest* (International Organization for Standard tion, 2003: Figure 4.5).

Kinds of Policies Required

What kinds of policies should be in place? The OAIS Reference Mo suggests that policies are required for archival storage (e.g., for manag migration), management (e.g., resource utilization, pricing), disas recovery, and security (e.g., physical access control).

Examples of policies developed for digital curation are available the PADI website (www.nla.gov.au/padi/topics/172.html). One these is the UK Data Archive's Preservation Policy. Section 5.3.1 no the policy for *Physical data preservation and storage* and illustrates h policies are written and what they contain:

> In order to best safeguard long-term preservation, the UKDA follows a policy of multiple copy resilience. Five versions of the complete preservation system are held: main near-line copy (on the main preservation server) and a shadow copy (on main preservation server). Both are held on the main area on the Hierarchical Storage Management (HSM) system and are presently accessed only by the dedicated preservation user. The storage media used for this copy is SDLT and disc cache area. The access online copy (on the mirror preservation server) is held in a RAID 5 disc system and copies are generated for user access and dissemination. There are also a near-site online copy kept on a RAID 5 disc system on a server located in another building within the University of Essex, and an off-site online copy. Finally a disc-based offline copy exists, which are held in either DVD-R or CD-R copy. The UKDA follows best practice in the storage and housing of magnetic and optical media. In particular, for environmental conditions for storage media (BS 4783, ISO/IEC22051, BS ISO 18921:2002 and BS ISO 18925:2002) and for the storage of archival materials (BS 5454). (Woollard, 2009: 10)

The Trusted Digital Repository certification places heavy emphasis policies (see Chapter 14). The requirements for achieving Trusted Dig Repository status state that policies must be in place. One of requirements is A3.2: "Repository has procedures and policies in pl and mechanisms for their review, update, and development as repository grows and as technology and community practice evol These policies need to be complete, readily available, and kept up date and evolve as circumstances change. They should address at l the core areas of "transfer requirements, submission, quality cont storage management, disaster planning, metadata management, acc rights management, preservation strategies, staffing, and securi These are accompanied by documented procedures about "day-to- practice and procedure" (RLG-NARA Task Force on Digital Reposit Certification, 2007: 13). Specifically noted in Appendix 3, which lists minimum documentation required (RLG-NARA Task Force on Dig Repository Certification, 2007: 81), are policies for legal permissi

policies and procedures relating to feedback, recording access actions and access, as well as documented procedures and disaster plans.

An essential guide to developing policy is *Policy-making for Research Data in Repositories* developed by the DISC-UK Data Share Project (Green, Macdonald, and Rice, 2009). Its valuable guidance covers six key areas where policies are helpful. The first of these areas relates to the content of the archive: its scope, the kinds of data it handles, file formats it will accept, and so on. The second area is policy relating to metadata: who has access to it, its reuse, types required, and sources and metadata schemas adopted. Third is policy relating to the ingest of data: who is eligible to deposit it, the quality requirements for that data, policy around confidentiality of data, and rights relating to that data. The fourth area is policy about access, use, and reuse of data. Fifth is policy applying to preservation of data: retention periods, fixity, and authenticity are included. Finally, policy about the withdrawal of data from the archive is important.

Although the importance of policy for digital curation is well recognized, policies are not as widespread as their significance warrants. Digital preservation policies were examined in a study commissioned by JISC in 2008. This study identified the lack of digital preservation policies and the resulting low priority paid to it in strategic planning. The two volumes of this study present an analysis of existing policies and offer excellent advice to developers of preservation policies through a model and numerous examples (Beagrie et al., 2008).

Costs of Curation

Planning for curation calls for estimates of the resources required and identification of where these resources will come from. Planning is difficult because the costs of curating data are not yet fully understood, despite considerable research into determining them and into the development of cost models.

Two British research projects that are investigating costs are the Life Cycle Information for E-Literature (LIFE) project and the Keeping Research Data Safe project. The LIFE project (www.life.ac.uk) has produced a model of the digital lifecycle and a methodology for estimating costs based on that lifecycle. The LIFE model establishes the costs associated with each phase of the lifecycle: creation or purchase; acquisition; ingest; metadata creation; bit stream preservation; content preservation; and access. It also acknowledges costs outside the lifecycle that need to be considered, such as costs of management and administration, systems and infrastructure, and economic adjustments such as inflation (Wheatley et al., 2007). A later phase of this project, under way in 2009–2010, is developing the model further by analyzing more examples and is producing a software tool for the prediction of costs.

The first phase of the Keeping Research Data Safe project investigated the costs of preserving research data in British universities (www.jisc.ac.uk/publications/documents/keepingresearchdatasafe.aspx). The second

phase (www.beagrie.com/jisc.php) builds on this work by identifying long-lived data sets and analyzing the costs of preserving them.

The costs of digital preservation are being investigated by the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (brtf.sdsc.edu). Economic sustainability is its focus—how to determine costs and then ensure that the resources are available for preservation activities by identifying sustainable economic models. Its interim report identified barriers to achieving economic sustainability. These included the inadequacy of current funding models that are often for one-off projects and not sustained, lack of recognition of the urgency of the issues, and fear that digital access and preservation are too hard to take on (Blue Ribbon Task Force, 2008).

Costs of ensuring accessibility to research data are also being actively investigated by other bodies. The Alliance for Permanent Access, whose aim is to "foster the development of an ecosystem of trusted digital repositories to enable Europe to fully exploit the potential of European scientific collaboration," made the theme of its 2008 conference *Keeping the Records of Science Accessible: Can We Afford It?* (Alliance for Permanent Access, 2008). The conference report noted that precise costing of digital preservation is not possible, as there are too many variables depending on such factors as the type, quality, and quantity of data and the access required to them. However, factors that influence costs can be determined. Also noted was the importance of timeliness as a factor influencing costs. It is far cheaper to properly curate data at the creation stage; adding or changing "bad" metadata is prohibitively expensive after time has passed. The experience of the Archaeology Data Service (ads.ahds.ac.uk), it was noted, is that the highest percentage of overall costs occur at the Acquisition and Ingest stages rather than the Storage and Preservations stages—42 percent and 23 percent, respectively (Alliance for Permanent Access, 2008: 3). Clearly, then, planning *before* curation activities begin is well worthwhile.

Even though the costs of digital curation are not yet fully known, practitioners are able to report costs in some areas from their experience. Costs of storage were the subject of recent postings on e-mail lists. These noted that costs of ingest and data management can dominate storage costs, and, as the quantities of data stored increase, the cost of power becomes a significant factor in overall costs. "And as the cost of purchasing storage goes down, the cost of managing it goes up" (Ashley, 2009).

The costs of preserving RAW file versus Tagged Image file Format (TIFF) files (both used to store data about images) were explored in a series of postings (e.g., Rusbridge, 2008). For image files, the compression used in the format is a factor; a JPEG 2000 file takes about one-third of the storage space as an uncompressed TIFF file. This makes a sizeable difference in costs of storage if large numbers of files are being considered. But the answer is not so simple. For example, because files formats act differently their choice has consequences for curation. The example was cited of a one-byte error having only a minimal effect on an uncompressed TIFF, but the same one-byte error could affect 17 percent of a JPEG 2000 file because of the way the file is constructed. Other factors noted

were the costs of labor, hardware maintenance, software maintenance, media replacement at specified intervals, capital equipment replacement, software licenses, and electricity. The frequency of data migrations also affects cost, as they involve intervention by people, so, no matter how much migration processes are automated, their labor costs are unlikely to decrease.

Ways to measure the costs of digital curation are not yet clearly established. It should not be forgotten, however, that many other factors besides costs play a role in determining the economic sustainability of digital curation. These include developing and maintaining secure business models, presenting compelling business cases, and collaborating in networks of preservation partners (LeFurgy, 2009: 425). All of these require considerable planning.

Summary: Planning for Active Management

The planning of digital curation activities occurs at every stage of the curation lifecycle. It should be based on policy, which also needs to be articulated for every stage of the curation lifecycle. The costs of curation should be considered when planning, but the current state of our understanding of costs makes such consideration problematic.

Planning of digital curation is not the prerogative of those who work with scientific or scholarly data. It is, of course, highly essential for libraries and archives that manage digital objects both for immediate access and for longer-term storage and later reuse, as the previous examples show. Planning is also highly applicable to many aspects of curating personal data. An obvious example is planning a regular backup regime for personal data—and implementing it!

The next chapter describes the fourth Full Lifecycle Action of the DCC Curation Lifecycle, *Community Watch and Participation*. It notes the reasons for and processes involved in keeping up-to-date and participating in developments to advance and improve curation activities.

PRESERVATION PLANNING: REVIEW

Key points: *Preservation Planning* is the ongoing process of planning data curation activities.

Key activities:

- Appreciate the need for planning at all stages of curation
- Develop plans for all stages of curation
- Periodically review and update curation procedures

References

Alliance for Permanent Access. 2008. *Keeping the Records of Science Accessible: Can We Afford It?* The Hague: Alliance for Permanent Access. Available: www.alliancepermanentaccess.eu/index.php?id=3 (accessed April 26, 2010).

Ashley, Kevin. 2009. "Cost of Digital Archiving." The dcc-associates Listserv, comment posted June 5, 2009. Available: www.mail-archive.com/dcc-associates@lists.ed.ac.uk/msg00175.html (accessed April 26, 2010).

Beagrie, Neil, Najla Semple, Peter Williams, and Richard Wright. 2008. *Digital Preservation Policies Study. Part 1, Final Report.* Salisbury: Charles Beagrie Ltd. Available: www.jisc.ac.uk/media/documents/programmes/preservation/jiscpolicy_p1finalreport.pdf (accessed April 26, 2010).

Blue Ribbon Task Force. 2008. *Sustaining the Digital Investment: Issues and Challenges of Economically Sustainable Digital Preservation.* Washington,

DC: Blue Ribbon Task Force. Available: brtf.sdsc.edu/biblio/BRTF_Interim_Report.pdf (accessed April 26, 2010).

British Atmospheric Data Centre. 2008. "DMP Template." Didcot, UK: BADC.

Clifton, Gerald. 2005. "Risk and the Preservation Management of Digital Collections." *International Preservation News* 36 (September): 21–23. Available: www.ifla.org/VI/4/news/ipnn36.pdf (accessed April 26, 2010).

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).

Donnelly, Martin. 2008. "PLANETS Testbed." Edinburgh: Digital Curation Centre (September 4, 2008). Available: www.dcc.ac.uk/resources/briefing-papers/technology-watch-papers/planets-testbed (accessed April 26, 2010).

Donnelly, Martin, and Sarah Jones. 2009. "Data Management Plan Content Checklist: Draft Template for Consultation." Edinburgh: Digital Curation Centre (June 17, 2009). Available: www.dcc.ac.uk/docs/templates/DMP_checklist.pdf (accessed April 26, 2010).

Green, Ann, Stuart Macdonald, and Robin Rice. 2009. *Policy-making for Research Data in Repositories: A Guide.* Version 1.2. Edinburgh: EDINA and University Data Library. Available: www.disc-uk.org/docs/guide.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).

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.

LeFurgy, William G. 2009. "NDIIPP Partner Perspectives on 'Economic Sustainability.'" *Library Trends* 57, no. 3 (Winter): 413–426.

RLG-NARA Task Force on Digital Repository Certification. 2007. *Trustworthy Repositories Audit & Certification: Criteria and Checklist.* Chicago: Center for Research Libraries. Available: www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf (accessed April 26, 2010).

Rusbridge, Chris. 2008. "Responses to RAW versus TIFF: Compression, Error and Cost-Related." Digital Curation Blog, comment posted July 2, 2008. Available: digitalcuration.blogspot.com/2008/07/responses-to-raw-versus-tiff.html (accessed April 26, 2010).

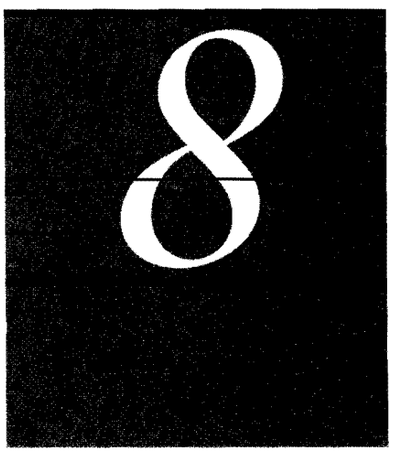
Serco Consulting. 2008. "UKRDS Interim Report." London: Serco Consulting (July 7, 2008). Available: ukrds.ac.uk/resources/download/id/17 (accessed April 26, 2010).

Wellcome Trust. "Q&A: Wellcome Trust Policy on Data Management and Sharing." London: Wellcome Trust. Available: www.wellcome.ac.uk/About-us/Policy/Spotlight-issues/Data-sharing/Data-management-and-sharing/WTX035045.htm (accessed April 26, 2010).

Wheatley, Paul, Paul Ayris, Richard Davies, Rory Mcleod, and Helen Shenton. 2007. *The LIFE Model.* v1.1. London: LIFE Project. Available: eprints.ucl.ac.uk/4831 (accessed April 26, 2010).

Woollard, Matthew. 2009. *UK Data Archive Preservation Policy.* Version 03.10. Colchester: UK Data Archive. Available: www.data-archive.ac.uk/news/publications/preservationpolicy.pdf (accessed April 26, 2010). Used by permission of the UK Data Archive, University of Essex.

Sharing Knowledge and Collaborating



This chapter investigates the Full Lifecycle Action *Community Watch and Participation*—the process of keeping up-to-date and participating in developments to improve and advance curation activities. In practice, this means that digital curators participate actively in collaborative activities. The key activities encompassed by Community Watch and Participation are:

- keeping up-to-date with digital curation activities and with developments in related areas,
- sharing data and participating in other activities underpinning data reuse,
- participating in the development of standards for digital curation, and
- participating in the development of tools and toolkits for digital curation.

IN THIS CHAPTER:

Keeping Up-to-Date

The "Community Watch" part of *Community Watch and Participation* refers to being fully aware of other activities in the digital curation community on an ongoing basis. There are many ways to maintain such an awareness.

Digital curation is a field about which a wealth of high-quality information is available on the web. It is also a field that changes rapidly and has few common understandings so far. These factors make it very important to keep up-to-date. Online materials may be the best source of information, but they vary in quality and currency. The quantity of printed material is increasing; at least four books on digital preservation were published between 2005 and 2007, and numerous journal articles are being published. These issues of quality and quantity have been recognized by organizations involved in digital curation and digital preservation. Consequently, there is now a curated database of

high-quality resources, and some websites also provide authoritative advice and guidance that has been reviewed and approved by knowledgeable practitioners. Some of these resources are listed later. The intention of this list, which is limited to English-language resources, is to indicate the types of resources that are available to digital curators, not to provide an authoritative comprehensive list of resources.

Starting Points

PADI: Preserving Access to Digital Information (www.nla.gov.au/padi/index.html), an essential starting point for all areas of digital curation, is a database established by the National Library of Australia. Its products include a quarterly newsletter about recent developments, *DPC/PADI: What's New in Digital Preservation* (www.dpconline.org/graphics/whatsnew/), produced in conjunction with the United Kingdom's Digital Preservation Coalition. Understanding the terminology used in digital curation is assisted by glossaries that include the Arts and Humanities Data Service's (2007) "AHDS Preservation Glossary" and the Society of American Archivists' "Glossary of Archival and Records Terminology" (Pearce-Moses, 2005). Authoritative guidance on specific topics is found in the *Curation Reference Manual* (Digital Curation Centre, accessed 2010). Produced by the Digital Curation Centre (DCC), this manual provides installments by experts in the field on a range of topics relevant to digital curation. The UNESCO (2003) *Guidelines for the Preservation of Digital Heritage* also provide authoritative advice. The Digital Preservation Coalition (2008) maintains the influential publication *Preservation Management of Digital Materials: A Handbook.*

Online Tutorials

Several online tutorials about digital preservation are available, although none of these covers digital curation in its entirety. Among them are the following:

- *Digital Preservation Management* (Cornell University Library, 2003–2007). This highly recommended tutorial is an essential introduction to digital preservation.
- The Canadian Heritage Information Network (CHIN) provides a number of tutorials related to creating and managing digital content (www.chin.gc.ca).
- The Joint Information Systems Committee (JISC) Digital Media's cross-media advice documents (www.jiscdigitalmedia.ac.uk/crossmedia) include "An Introduction to Digital Preservation" (accessed 2010) and "Establishing a Digital Preservation Strategy" (accessed 2010).

There are many tutorials on specific technical topics on the web, for example, *OAI for Beginners: The Open Archives Forum Online Tutorial* (Open Archives Forum, 2003).

Project Websites

The websites of digital curation and digital preservation projects are a fruitful source of useful material. These must be used with care, as the information they present is not always current. Project funding may have ended and work ceased, but the website may still be available.

- Cultural, Artistic, and Scientific Knowledge for Preservation, Access, and Retrieval (CASPAR; www.casparpreserves.eu). CASPAR recently added Training Lectures, "a collection of videos of talks and screen captures of software—all about digital preservation."
- Digital Preservation Coalition (DPC; www.dpconline.org/graphics/index.html). This is not a project but is included here because its website contains useful reports, including a series of Technology Watch reports.
- Digital Preservation Europe (DPE; www.digitalpreservationeurope.eu). Links to reports, briefing papers, and position papers are in the "DPE Publications" pages of the website.
- Electronic Resource Preservation and Access Network (ERPANET; (www.erpanet.org). This project ended in 2007 but during its existence produced a wide range of materials that are still useful, although some are now outdated.
- National Digital Information Infrastructure & Preservation Program (NDIIPP; (www.digitalpreservation.gov). This is "a Collaborative Initiative of the Library of Congress."
- Paradigm Project (www.paradigm.ac.uk). This project, funded from 2005 to 2007, produced a workbook that covers many aspects of curation.
- PLANETS (www.planets-project.eu) provides reports on aspects of its toolkit development.
- Sustaining Heritage Access through Multivalent ArchiviNg (SHAMAN) (shaman-ip.eu/shaman). This project is aimed at developing a long-term digital preservation framework.

Blogs and E-mail Lists

Among blogs that are specifically about digital curation are the Digital Curation Blog (digitalcuration.blogspot.com), "inspired by the Digital Curation Centre to discuss issues relating to the curation and long term preservation of digital science and research data," and the DCC Blawg (dccblawg.blogspot.com), which focused on the legal aspects of digital curation. Postings about digital curation appear regularly in many other blogs. E-mail lists whose primary interest is digital curation include DIGITAL-PRESERVATION List (www.jiscmail.ac.uk/cgi-bin/webadmin?A0=digital-preservation), operated by JISC, a major U.K. funding body for higher education that has had a major role in promoting digital preservation, and padiforum-l (www.nla.gov.au/padi/forum), owned

by the National Library of Australia and affiliated with the PADI web portal for digital preservation resources. PADI identifies other discussion lists (www.nla.gov.au/padi/format/list.html) relevant to digital curation.

Online Journals

Articles about digital curation appear in a wide range of journals and can be located using indexing and abstracting databases available through libraries. Journals that regularly publish articles about digital curation include *Ariadne* (www.ariadne.ac.uk), which notes current digital library initiatives and technological developments; *D-Lib Magazine* (www.dlib.org), which focuses on "digital library research and development, including . . . new technologies, applications, and contextual social and economic issues"; and the *International Journal of Digital Curation* (www.ijdc.net/index.php/ijdc/issue/current), published by the DCC. The DCC website provides a list of curation- and preservation-related journals (www.dcc.ac.uk/IJDC). PADI also provides a list of journals and newsletters (www.nla.gov.au/padi/format/journal.html).

Other Sources

The DPE website provides an annotated list of online resources (www.digitalpreservationeurope.eu/registries/resources) of relevance and importance to a wide range of digital curation activities. Discipline-specific websites also contain relevant resources. An example is the list of digital curation resources (nssdc.gsfc.nasa.gov/nost/curation.html) available on the NASA/Science Office of Standards and Technology (NOST) website. The number of training opportunities in digital curation is increasing. One of DPE's aims is to coordinate training opportunities, and to this end it maintains a Registry of Digital Preservation Trainers (www.digitalpreservationeurope.eu/registries/trainers) and an events register (www.digitalpreservationeurope.eu/events/events).The DCC also maintains lists of forthcoming events (www.dcc.ac.uk/events) that can help locate training opportunities.

Collaboration: Intrinsic to Digital Curation

The "Participation" part of *Community Watch and Participation* refers to the need for collaboration in the digital curation community. Collaboration is one of the keys to effective curation. All communities involved in curation—data creators, users, and in fact all stakeholders—should participate in discussions about the challenges posed and in creating helpful responses to these challenges. Collaborative efforts are, in fact, "more the norm than the exception" (Jordan et al., 2008: 6).

Collaboration is, in fact, firmly embedded in digital curation practice. Active management of data for current and future use relies on effective sharing of data, which, in turn, relies on agreement on and adoption of standards. Partnerships have been important in digital curation from its

inception, because it was quickly realized that no single organization could adequately archive, preserve, and provide access to digital materials. The reasons for this include the scale of digital curation issues, uncertainty about how to address them, and the high cost of digital curation in relation to the financial resources available. Collaboration ensures the best use of resources through sharing expertise and experience and through developing and building technical resources and solutions that can be shared.

The benefits of collaboration have been rehearsed in many publications and are well understood. The UNESCO (2003) *Guidelines for the Preservation of Digital Heritage* identify many of them, specifically in relation to digital curation. These include access to a wider range of expertise; sharing of the costs of developing software and systems; access to tools and systems of other organizations; the sharing of learning opportunities; encouragement of influential stakeholders to take digital curation seriously; increased ability to influence data producers and system developers; joint research and development of standards and practices; and enhanced ability to attract resources and other support for well-coordinated curation programs at regional, national or sectoral levels (UNESCO, 2003: 64–65).

Examples of collaboration in digital curation are easy to find. Websites such as PADI and that of the DCC, noted in "Keeping Up-to-Date" earlier in this chapter, indicate one mechanism by which expertise is shared. International conferences such as the DCC's series of International Digital Curation Conferences also provide forums where expertise can be shared. The DPE project was established specifically to share European expertise in the field of digital curation by "pooling of the complementary expertise that exists across the academic research, cultural, public administration and industry sectors in Europe" (Digital Preservation Europe, "DPE: Digital Preservation Europe," accessed 2010). Both the DCC and DPE have as stated aims the sharing of learning opportunities, with the DPE, for example, running the "Digital Preservation Exchange Programme" (Digital Preservation Europe, accessed 2010).

The DCC's series of international conferences on digital curation are not the only conferences where expertise is shared. Others held on a regular basis include the International Conference on Digital Preservation (iPres). Most conferences of professional library and archives organizations now include sessions about digital curation, such as the conferences held by national bodies (the Society of American Archivists and the American Library Association) and by regional organizations. Many conferences that focus on digital libraries, such as the Joint Conference on Digital Libraries (JCDL), the European Conference on Digital Libraries (ECDL) and the International Conference on Asian Digital Libraries (ICADL), also include sessions or streams devoted to digital curation.

Sharing of the costs of developing software and systems is practiced through the adoption of open-source concepts. For example, two of the key digital preservation repository systems, Fedora (www.fedora.info) and DSpace (www.dspace.org), are open source and have a large international community of participants who contribute to their development.

AN EXAMPLE OF COLLABORATION

Some material quoted in this book has a Creative Commons Non-Commercial (CC-NC) license. This means it can be used without permission for noncommercial purposes, and readers may reuse it, provided they attribute the original source in the manner specified by the author or licensor. (The Creative Commons website provides more information: creativecommons.org.)

This willingness to share published information about digital curation illustrates the international and collaborative nature of digital curation. Developments and practice in one region are keenly observed and adopted and modified to suit local requirements in other regions. Many digital curation projects funded by public money are required to make documentation of their activities freely available.

Persistent lobbying of influential stakeholders to take digital curation seriously is carried out by groups such as the DPC and the DCC in the United Kingdom, both of which have lobbying as a stated objective, and the NDIIPP in the United States, also with a strong lobbying role. Joint research and development is evident in the development of key standards used in digital curation. An example is METS, whose Editorial Board has members from Germany, the United Kingdom, and the United States (www.loc.gov/standards/mets/mets-board.html); another is the development of the Trusted Digital Repository concept (see Chapter 14) by a working group with members from the United Kingdom, Germany, the United States, France, and Australia. These two examples are by no means unusual in digital curation.

Projects that aim to make data sharing easier also demonstrate high levels of collaboration. The sharing of expertise is fully demonstrated in the DISC-UK DataShare project (www.disc-uk.org/datashare.html), which ran from 2007 to 2009. It involved the collaboration of partners to develop new models, workflows, and tools for sharing research data sets.

Projects that share archival storage resources, such as MetaArchive, also demonstrate collaboration. The MetaArchive Cooperative (www.metaarchive.org), established in 2003 with funding from the Library of Congress's NDIIPP, is a coalition of universities and research libraries that provides "low-cost, high-impact preservation services to help ensure the long-term accessibility of the digital assets of universities, libraries, museums, and other cultural heritage institutions" (MetaArchive Services Group, accessed 2010). It is based on the open-source Lots of Copies Keep Stuff Safe (LOCKSS) software, which allows digital preservation to be carried out collaboratively at a series of geographically distributed sites. In addition to providing an affordable service to its members, MetaArchive assists other groups to create similar collaborative digital preservation networks, hosts activities to support and train groups wishing to establish networks, and fosters awareness of digital preservation issues (Halbert and Skinner, 2008: 5).

Standards: Essential for Digital Curation

Effective digital curation requires the development and implementation of standards. (Chapter 3 noted one important standard, the OAIS Reference Model, formalized as ISO standard 14721:2003, which is widely used as the basis of planning digital archives.) A fundamental aspect of digital curation is the sharing and reuse of data. This implies interoperability of systems—the ability of software and hardware to exchange and use information. For reliable and consistent interoperability, standards are essential. They are the basis upon which digital curation systems that work are built. Standards, however, require consensus among all who apply them so that confusion and misunderstanding are reduced; achieving consensus, in turn, requires knowledge of what is happening in digital curation—in other words, maintaining a community watch.

Standards apply to all digital curation activities: data capture; citation; annotation; classification; achieving interoperability; software integration; representation information—the list is long. The chapters in Part III note which standards are relevant to each activity of the curation lifecycle.

The first step is to identify (or develop if they don't already exist) the standards that are relevant for curation activities. Many standards for digital curation were developed by the DIFFUSE project, funded by the European Union (EU), which was concerned with standards relevant to the information society. The DCC continues to maintain and update the registry of these standards (Digital Curation Centre, "DCC DIFFUSE Standards Frameworks," accessed 2010). It can be searched through the DCC Curation Lifecycle actions. Standards have not yet been developed for all aspects of digital curation, and digital curators may find themselves involved in developing new standards.

Three examples illustrate the significance of standards for digital curation and of participating in standards development.

A standard for exporting data: The instrumentation used in creating data in research experiments is typically commercially manufactured. Many of these instruments use data-capture tools specific to a manufacturer, which are based on data standards developed by that manufacturer. There is little uniformity, even for similar types of instrumentation. The scientific community is increasing pressure on manufacturers to conform to standards, at least in the way that data that they create and collect can be exported from the software used for creation and collection (e.g., enabling XML output) so that data from a wide range of instruments can be combined.

A standard for stable data formats: Using standard data formats that will remain accessible over time is a commonly applied digital preservation strategy. Standards that are stable and have been widely adopted are much more likely to be supported over a long period. Standards that are open (i.e., not proprietary) are less likely to become obsolete in a short period, because there is a large user base willing to participate in ensuring that the standards are maintained. XML is often adopted, as it is a stable standard with a very good track record, is in increasingly widespread use, and has a large user base. Chapter 10 notes in more detail the requirements for stable data formats.

A standard for sharing data: One requirement for interoperability is that data and their associated metadata are expressed in a standard format so that different software can recognize them and then act on them. For metadata, METS (Library of Congress, accessed 2010) provides a standard for encoding metadata for digital objects. It is widely used in digital curation environments.

Tools and Toolkits

Digital curation processes are often characterized as "artisan" or "hand-crafted," referring to their labor-intensive nature. This limits considerably the quantities of data that can be curated using these processes. Automation

of curation workflows and processes is commonly understood to be essential for improving digital curation, because it increases the ability to curate larger quantities of data and reduces the costs of doing so. Automated workflows and procedures need software tools—tools that can be applied in many areas of digital curation. Some curation actions where software tools can be applied include the following:

- Identification of digital objects (e.g., where they are located, what formats they are in)
- Describing digital objects (e.g., automated metadata creation)
- Manipulating data (e.g., data management, data storage, repositories)
- Preserving data (e.g., migration)
- Rights management and access control (e.g., restricting access to authorized users of a system)

Tools are being developed by many digital curation research and development projects. Recent EU-funded projects such as CASPAR and PLANETS are developing tools to advance automation in digital curation. CASPAR (www.casparpreserves.eu) aims to "produce tools and techniques to support digital preservation and make it easier to share the cost" (Giaretta, 2007). These tools must be useful and usable, easy to use, requiring little effort to adopt, sustainable after the CASPAR project ends, and open source. PLANETS (www.planets-project.eu) has developed prototype tools and services for preservation planning, preservation action, and preservation characterization. It has made available Plato, a preservation planning workflow tool based on a toolkit of other preservation planning tools, such as PLANETS-compliant migration tools for digital objects, emulation tools for specific environments, and characterization tools that extract significant properties from digital objects. In the United States, NDIIPP provides a long list of tools and services designed, developed, or used by NDIIPP partners (www.digitalpreservation.gov/partners/resources/tools/index.html). These tools cover most curation activities.

If they are to be adopted widely, the tools developed for digital curation must be *usable* and *useful*. These characteristics and tools that are in common use are described in more detail in Chapter 13. Community watch activities monitor the development of curation tools on an ongoing basis to determine whether new tools are applicable to local data curation activities. In addition, input into tool development, where feasible, as part of collaborative approaches to digital curation will ensure that they are usable and useful.

Summary: Collaboration Is the Key

Engaging with the wider digital curation community has many positive benefits. The first step is to become aware of current thinking and practice in the field so that local practice develops as best practice. For example,

monitoring changes in technology indicates when hardware and software are in danger of becoming obsolete. (This is known as *technology watch.*) Participating in some of the many collaborative activities that characterize digital curation and in the development of standards in the field provides more effective outcomes.

The next chapter is the first in Part III, "The Digital Curation Lifecycle in Action." This section of the book examines the DCC Curation Lifecycle's Sequential Actions. First, Chapter 10 notes the development and planning of data creation procedures with digital curation activities and outcomes in mind.

COMMUNITY WATCH AND PARTICIPATION: REVIEW

Key points: *Community Watch and Participation* is the ongoing process of keeping up-to-date with data curation activities and developments in related areas, and participating in developments to advance and improve curation activities.

Key activities:

- Keep up-to-date with data curation activities and developments in related areas
- Share data and participate in activities underpinning data reuse
- Participate in standards development
- Participate in the development of tools and toolkits for data curation

References

Arts and Humanities Data Service. 2007. "AHDS Preservation Glossary." London: Arts and Humanities Data Service (October 17, 2007). Available: ahds.ac.uk/exec/creating/glossary.htm (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).

Digital Curation Centre. *Curation Reference Manual.* Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/resources/curation-reference-manual (accessed April 26, 2010).

———. "DCC DIFFUSE Standards Frameworks." Edinburgh: Digital Curation Centre. Available: www.dcc.ac.uk/resources/standards/diffuse (accessed April 26, 2010).

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

Digital Preservation Europe. "Digital Preservation Exchange Programme (DPEX)." Glasgow: DPE. Available: www.digitalpreservationeurope.eu/exchange (accessed April 26, 2010).

———. "DPE: Digital Preservation Europe." Glasgow: DPE. Available: www.digitalpreservationeurope.eu (accessed April 26, 2010).

Giaretta, David. 2007. "The CASPAR View on What Digital Curators Do and What They Need to Know." Paper presented at DigCCurr 2007, Chapel Hill, NC, April 18–20, 2007. Available: www.casparpreserves.eu/Members/cclrc/Presentations/the-caspar-view-on-what-digital-curators-do-and-what-they-need-to-know-research-perspectives/at_download/file (accessed April 26, 2010).

Halbert, Martin, and Katherine Skinner. 2008. "The MetaArchive Cooperative: A New Collaborative Service Organization Providing a Distributed Digital Preservation Infrastructure." *CLIR Issues* no. 66. Available: www.clir.org/pubs/issues/issues66.html (accessed April 26, 2010).

Joint Information Systems Committee. "Establishing a Digital Preservation Policy." Bristol: JISC Digital Media. Available: www.jiscdigitalmedia.ac.uk/crossmedia/advice/establishing-a-digital-preservation-policy (accessed April 26, 2010).

———. "An Introduction to Digital Preservation." Bristol: JISC Digital Media. Available: www.jiscdigitalmedia.ac.uk/crossmedia/advice/an-introduction-to-digital-preservation (accessed April 26, 2010).

Jordan, Christopher, Ardys Kozbial, David Minor, and Robert H. McDonal "Encouraging Cyberinfrastructure Collaboration for Digital Preservation Paper presented at iPres 2008, British Library, London, September 3(2008. Available: www.bl.uk/ipres2008/presentations_day2/39_Jordan.p (accessed April 26, 2010).

Library of Congress. "METS: Metadata Encoding & Transmission Standard. Washington: Library of Congress. Available: www.loc.gov/standards/me (accessed April 26, 2010).

MetaArchive Services Group. "About MetaArchive." Atlanta: MetaArchiv Services Group. Available: www.metaarchive.org/about (accessed April 2(2010).

Open Archives Forum. 2003. *OAI for Beginners: The Open Archives Forun Online Tutorial.* Bath: OA-Forum and UKOLN. Available: www.oaforun .org/tutorial (accessed April 26, 2010).

Pearce-Moses, Richard. 2005. "Glossary of Archival and Records Terminology." Chicago: Society of American Archivists (2005). Available: www.archivist .org/glossary (accessed April 26, 2010).

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