ISSN 1747-1524

DCC | Digital Curation Manual

Instalment on "Appraisal and Selection"

http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection/

Ross Harvey
School of Information Studies
Charles Sturt University
http://www.csu.edu.au/faculty/sis/

January 2007
Version 1.0

Legal Notices

Catalogue Entry

Title	DCC Digital Curation Manual Instalment on Appraisal and Selection
Creator	Ross Harvey (author)
Subject	Information Technology; Science; Technology--Philosophy; Computer Science; Digital Preservation; Digital Records; Science and the Humanities.
Description	Instalment on the role of selection and appraisal within the digital curation life-cycle. Describes the increasingly important role of selection and appraisal for digital curation, some practical applications, the topic's place within the OAIS reference model, and advice on developing institution-specific selection frameworks.
Publisher	HATII, University of Glasgow; University of Edinburgh; UKOLN, University of Bath; Council for the Central Laboratory of the Research Councils.
Contributor	Seamus Ross (editor)
Contributor	Michael Day (editor)
Date	10 June 2006 (creation)
Type	Text
Format	Adobe Portable Document Format v.1.3
Resource Identifier	ISSN 1747-1524
Language	English
Rights	© HATII, University of Glasgow

Citation Guidelines

Ross Harvey, (June 2006), "Selection and Appraisal", *DCC Digital Curation Manual*, S.Ross, M.Day (eds), Retrieved <date>, from
http://www.dcc.ac.uk/resource/curation-manual/chapters/appraisal-and-selection

About the DCC

The JISC-funded Digital Curation Centre (DCC) provides a focus on research into digital curation expertise and best practice for the storage, management and preservation of digital information to enable its use and re-use over time. The project represents a collaboration between the University of Edinburgh, the University of Glasgow through HATII, UKOLN at the University of Bath, and the Council of the Central Laboratory of the Research Councils (CCLRC). The DCC relies heavily on active participation and feedback from all stakeholder communities. For more information, please visit www.dcc.ac.uk. The DCC is not itself a data repository, nor does it attempt to impose policies and practices of one branch of scholarship upon another. Rather, based on insight from a vibrant research programme that addresses wider issues of data curation and long-term preservation, it will develop and offer programmes of outreach and practical services to assist those who face digital curation challenges. It also seeks to complement and contribute towards the efforts of related organisations, rather than duplicate services.

DCC - Digital Curation Manual

Editors

Seamus Ross
Director, HATII, University of Glasgow (UK)
Michael Day
Research Officer, UKOLN, University of Bath (UK)

Peer Review Board

Preface

The Digital Curation Centre (DCC) develops and shares expertise in digital curation and makes accessible best practices in the creation, management, and preservation of digital information to enable its use and re-use over time. Among its key objectives is the development and maintenance of a world-class digital curation manual. The *DCC Digital Curation Manual* is a community-driven resource—from the selection of topics for inclusion through to peer review. The Manual is accessible from the DCC web site (http://www.dcc.ac.uk/resource/curation-manual).

Each of the sections of the *DCC Digital Curation Manual* has been designed for use in conjunction with *DCC Briefing Papers*. The briefing papers offer a high-level introduction to a specific topic; they are intended for use by senior managers. The *DCC Digital Curation Manual* instalments provide detailed and practical information aimed at digital curation practitioners. They are designed to assist data creators, curators and re-users to better understand and address the challenges they face and to fulfil the roles they play in creating, managing, and preserving digital information over time. Each instalment will place the topic on which it is focused in the context of digital curation by providing an introduction to the subject, case studies, and guidelines for best practice(s). A full list of areas that the curation manual aims to cover can be found at the DCC web site (http://www.dcc.ac.uk/resource/curation-manual/chapters). To ensure that this manual reflects new developments, discoveries, and emerging practices authors will have a chance to update their contributions annually. Initially, we anticipate that the manual will be composed of forty instalments, but as new topics emerge and older topics require more detailed coverage more might be added to the work.

To ensure that the Manual is of the highest quality, the DCC has assembled a peer review panel including a wide range of international experts in the field of digital curation to review each of its instalments and to identify newer areas that should be covered. The current membership of the Peer Review Panel is provided at the beginning of this document.

The DCC actively seeks suggestions for new topics and suggestions or feedback on completed Curation Manual instalments. Both may be sent to the editors of the *DCC Digital Curation Manual* at curation.manual@dcc.ac.uk.

Seamus Ross & Michael Day.
18 April 2005

Biography

Ross Harvey is Professor of Library and Information Management at Charles Sturt University, NSW, Australia. He has worked as a librarian, academic and consultant in New Zealand, Australia, and Singapore. His interest in preservation stems from a background in historical bibliography coupled with an early position as the National Library of New Zealand's Newspaper Librarian. His publications about preservation includes several books: *Preservation in Libraries: Principles, Strategies and Practices for Librarians* and *Preservation in Libraries: A Reader*, both published in 1993 by Bowker-Saur and, in 2005, *Preserving Digital Materials*, published by K.G. Saur. Harvey's interest in archives management has led him to investigate practice in the recordkeeping sectors, and he attempts in his research to explore the intersections and commonalities between recordkeeping and library practices. He has recently explored archival appraisal theory and practice to see whether it might have lessons that could improve data curation processes. He has also recently investigated Australian digital preservation practice to see if it, too, has lessons that could assist data curators. A full CV is available on the web at http://www.elibank.net.

Table of Contents

Introduction and Scope 8
Need for Appraisal and Selection 8
Background and Developments to Date 9
Definitions 9
Professional Responsibilities 9
Traditional Appraisal and Selection Concepts 10
Modifying Appraisal and Selection Criteria for Data 10
Need for Viable Data Appraisal and Selection Strategies 11
Early Data Appraisal and Selection Frameworks 11
How does Appraisal and Selection Apply to Digital Curation? 12
The Data Life-Cycle 12
Context and Community 12
Data Creation and Data Creators 13
Increased Role of Stakeholders 13
Discipline-based Approaches to Selection 14
Technical Capacity to Preserve as a Selection Factor 14
Legal Issues as a Selection Factor 14
Data Curation and Data Curators 15
The Influence of Copyright Legislation on Appraisal and Selection 15
Metadata's Role in Appraisal and Selection 15
Data Re-use and Re-users 16
Exactly What Are we Preserving?: Essential Elements 16
Selecting Essential Elements 16
Re-appraisal 17
A Generic Appraisal and Selection Framework 17
Guidance from the Literature 17
Risk Management 18
The Generic Framework 18
Appraisal and Selection in Action 19
Scarcity of Data Appraisal and Selection Policies 19
Selection Policies and Criteria: Libraries and Digital Libraries 19
SunSITE Digital Library 19
Library of Congress 20
University Libraries 20
Genre Selection Policies 21
E-repositories 22
Appraisal Policies and Criteria: Recordkeeping 23
Recordkeeping Principles 23
Non-specific Nature of Appraisal Policies 23
Macro-appraisal 23
Appraisal and Selection Policies and Criteria: Research Data 24
Research Data: A Special Case? 24
Examples of Selection Criteria for Research Data 24
The State Library of Victoria, Melbourne, Australia 25

Next steps 26
Developing an Institution-specific Selection Framework..... 26
Key Questions and Actions..... 26
Content, Structure and Context..... 28
Future developments 28
Conclusion 29
References……32
Appendices……37
Annotated List of Key External Resources……...39

Introduction and Scope

Need for Appraisal and Selection

The ever-increasing quantities of data being produced in digital form (readily demonstrated by the growth rate of the Internet – see www.isc.org/ds/), the rapidity of change of computer and information technology, and the changing ways in which data is produced and becomes available to its communities of users, combine to present us with new and complex challenges in data curation. (The term *data* is used in this chapter to mean material in digital form.) It is unlikely that those responsible for the maintenance and accessibility of data over time – data curators – will be able to preserve all data for which they are responsible. Resources are finite, so tools are required to assist data curators to appraise and select which data is most important to maintain and provide access to, both in the short term as active data, and the longer term for re-use. These tools are essential for responsible, effective data curation practice. The guidance offered by such appraisal and selection tools becomes increasingly necessary as the rate of data production continues to outstrip the rate at which resources become available for data curation.

Key Assumptions and Challenges

Current thinking about selecting data for preservation can be characterised as based around two ways of thinking which are polar opposites. One way subscribes to a technological deterministic future in which computer storage costs reduce and processing power increases, allowing us to keep all data and avoiding the need to make decisions about what we should archive and maintain for future use. This approach is limited just to bit preservation and therefore considered not appropriate for knowledge preservation. The other way, on which this chapter is based, makes some key assumptions:

- It is not practical to maintain access to all data indefinitely
- It is not desirable to maintain access to all data indefinitely
- Appraisal and selection of data for preservation is based on their significance and continuing value
- We need to select and preserve more than just the data themselves, in order to understand them in the future
- An appraisal or selection policy that clearly sets out the processes and the basis for making selection decisions is necessary
- Appraisal and selection decisions should be consistently applied
- Appraisal and selection decisions must be based on a clear understanding of the objectives of the organization that accepts preservation responsibility.

These assumptions raise many questions. Some examples: How do data curators determine what is most important? What is *significant*? Significant to whom? Significant for what reason? How might we determine *continuing value*? Value to whom? When – in ten years time, twenty, one hundred? Some guidance with answering these and other questions comes from examining the practices of recordkeepers (this term refers to archivists and records managers) and

librarians who, in order to meet their responsibilities for preserving documents for future use, have developed criteria and processes for identifying the documents to which they will devote resources to ensure their preservation. These, however, have been developed for collections that are primarily paper-based. They do not automatically translate to data, and need to be modified to ensure that they can be applied effectively to digital materials. For example, selection decisions for digital materials necessarily have heavy ongoing resource implications, unlike those for paper-based material for which the ongoing costs of maintenance can be suspended for periods of time without major detrimental effect (the 'benign neglect' concept). Vogt-O'Connor reminds us that 'program costs don't cease when the Web site disappears' (Vogt-O'Connor 2000).

This chapter investigates how appraisal and selection processes have been applied to data preservation, provides some guidelines to assist data curators in making appraisal and selection decisions, and indicates some directions in which further investigation is needed. It specifically notes the role of stakeholders, existing selection guidelines and policies, and areas in which new policies need to be developed.

Background and Developments to Date

Appraisal and selection are crucial to data preservation because of the fact that it is not feasible in resource terms, and perhaps not desirable, to maintain access to all data indefinitely. Data curators need to make decisions about what to keep and provide access to into the future. This simple statement poses many challenges, including questions of definition (for example, of *significance*, *continuing value*), scope (what do we need to select in addition to the data?) and process (what needs to be decided when?). This section notes some appraisal and selection practices, considers why they cannot be applied without alteration to data, and indicates some early attempts to articulate appraisal and selection criteria for data.

Definitions

Some preliminary working definitions are needed. *Appraisal* is a concept familiar to archivists: 'the process of evaluating records to determine which are to be retained as archives, which are to be kept for specified periods and which are to be destroyed' (Ellis, 1993, p.461). *Selection* is the process of deciding what will be added to a library's collection.

Professional Responsibilities

Librarians and archivists have developed appraisal and selection techniques to meet one of their responsibilities: to preserve significant documents for future use. This responsibility, core business of many institutions, has been emphasised by archivists, who have developed a considerable body of theory and practice about appraisal. This has, however, been developed primarily for paper-based collections, and cannot be applied to digital materials without modification and revision. Selection and appraisal processes are inherently value-laden. For example, because records managers select records for retention based on 'risk avoidance, market opportunities, or desires to avoid embarrassment or accountability', the outcome 'inevitably will privilege the needs of business or government in terms

of the issues that get addressed, the allocation of resources, and the long-term survival of records' (Cook, 2000, p.8). The records that survive will not reflect the full range of human experience, but rather the concerns of administrators. We need to remember that 'Every choice to preserve is at the expense of something else' (UNESCO, 2003, p.73). Responsible selection and appraisal practice must acknowledge try to minimise such biases.

Traditional Appraisal and Selection Concepts

There is considerable consensus about the criteria used in appraisal and selection of non-digital materials. Archival appraisal practice uses the concept of *archival value*, assessed by considering *administrative value* (usefulness for the conduct of business), *fiscal value* (usefulness for financial business), *legal value* (worth for conduct of legal business), intrinsic value (inherent nature and artefactual significance), *evidential value* (value as record of the record creator's origins, functions and activities), and *informational value* (usefulness of content for more general research purposes) (Tibbo, 2003, pp.29-30). Library practice for selection of materials for long-term preservation focuses on maintaining physical items in their original formats and applies five key criteria: *evidential value*, *aesthetic value*, *market value*, *associational value*, and *exhibition value*. Additional criteria may be applied, such as physical condition, resources available, use, social significance. An important difference between archival appraisal practice and library selection practice is the emphasis on context (see 4.2) in archival appraisal.

Modifying Appraisal and Selection Criteria for Data

Appraisal and selection criteria and processes developed for traditional (usually paper-based) materials cannot be applied to data without modification. Different emphases are required. For example, there needs to be greater emphasis on our technical capability to preserve data, magnifying practical considerations beyond those associated with preserving paper materials where taking no preservation action was not necessarily harmful. The ongoing costs of maintaining data (the 'digital mortgage') require greater emphasis. A third difference is the need to make preservation decisions early in the existence of data, otherwise they will become inaccessible or disappear. The UNESCO *Guidelines for the Preservation of Digital Heritage* remind us that 'it may not be possible to wait for evidence of enduring value to emerge before making selection decisions' (UNESCO, 2003, p.74). Other differences include the challenges posed by new digital genres, the difficulties associated with deciding precisely which attributes of data should be preserved, and the legal complexities associated with determining ownership of intellectual property rights. Effective criteria and processes for appraisal and selection of data for preservation must take account of these differences.

Digital is different, and this inescapable fact inevitably affects selection and appraisal decisions. To begin with, there is more of it, 'more things – more information, more records, more publications, more data – than we have the means to keep' (UNESCO, 2003, p.73). The networked environment has removed many of the pre-digital environment's quality control mechanisms, so that data

curators may need to cope with decreasing (or at least variable) data quality, in addition to increased quantity. This places, for some organizations, an even greater emphasis on selection. And, perhaps most critically, unlike paper-based and other traditional artefacts, there is no 'comfort zone' (Jones & Beagrie, 2001, p.30) during which selection decisions can be made before materials deteriorate; the time frame for data is much shorter.

Need for Viable Data Appraisal and Selection Strategies

At present we lack scalable and defensible appraisal and selection strategies for digital preservation. Most selection scenarios are unsatisfactory in some way. For example, the strategy of 'picking the low-hanging fruit' (preserving what is easiest to preserve) is unsatisfactory because what is easiest to preserve is unlikely to be of particular value; letting the marketplace decide links definitions of value to commercial considerations, rather than to long-term societal or organisational value (Burrows, 2000, p.148). We can find some guidance with developing viable strategies in the general principles on which library selection criteria and archives appraisal theory and practice are based. However, there is increasing evidence that generalized selection criteria are not appropriate for data, and that sectoral differences should be further investigated and appraisal and selection strategies developed on this basis.

Early Data Appraisal and Selection Frameworks

Some early thinking about selection of data for preservation is found in the Cedars Project Team report (Cedars Project Team 2002), in its concept of a digital object's *significant properties*, 'the level of content and functionality retained', which are derived in the context of specific user communities and an organisation's preservation responsibilities. Intellectual property rights are strongly emphasised in the Cedars report and their negotiation is the first preservation requirement. The report concludes that selection should be based on the 'estimated value of the material, the cost of storage and support mechanisms, and the production of metadata to support the material' (Cedars Project Team, 2002, p.53).

Also informative is the *Decision Tree for Selection of Digital Materials for Long-term Retention*, in the Digital Preservation Coalition's handbook (Jones & Beagrie, 2001, section 4). This provides four groups of questions, about:

- *Policy*: does the institution have a selection policy? Does the material fit into it?
- *Legal and intellectual property issues*: have, and can, acceptable rights be negotiated?
- *Technical* issues: can the file format be handled, currently and in future? Is transfer to a more manageable format format?
- The existence of *documentation, ancillary data and metadata*: is there sufficient?
- There are currently no commonly-accepted standards for appraisal and selection of digital materials. A summary of much of the experience to date from the digital heritage sector can

be found in the UNESCO *Guidelines for the Preservation of Digital Heritage* (UNESCO, 2003, chapter 12).

How does Appraisal and Selection Apply to Digital Curation?

The Data Life-Cycle

Viable appraisal and selection criteria for digital materials need to take account of factors which assume greater importance than for non-digital materials, such as **intellectual property rights**, and the **need to preserve more contextual information** about the materials. Stakeholders play an increased role in this process. Although generic appraisal and selection frameworks for data are helpful, they need to be significantly modified for specified communities and for specific categories of material. In particular, *context* must be taken account of: the context in which data were created, in which they are used, and in which they need to be maintained in the future. For example, prioritising of archival materials is based on the archive's statutory obligations and business objectives: in a sound broadcast archive, the re-use in program production is a key factor to be accounted for in appraisal decisions; a national sound archive has legal deposit responsibilities that affect decisions about what to maintain; and a research archive will consider the needs of researchers.

Informing these decisions is the data life-cycle model – data is created, maintained for active use, archived, preserved, then accessed and re-used, and disposed of or transferred to another custodian. This has far-reaching implications for data curation. Significant data need to be identified, appraisal and selection decisions need to be made, the critical aspects of those data (such as the attributes that determine their authenticity) need to be determined, and management and preservation decisions that will ensure ongoing access must be made right from the start of the data life-cycle.

Context and Community

As noted above, a universally-applicable appraisal and selection framework for data is not realistic. Different kinds of digital materials, created in different contexts for different stakeholders, require different approaches to appraisal and selection. For instance, the conditions under which data are acquired by a national library as part of legal deposit legislation are different from scientific datasets, and they both have different characteristics from records of business transactions created in digital form, and so on. What assists us here to develop viable appraisal and selection frameworks is the concept of *context.*

Every data curation programme is working within a specific context. The context for national libraries is the nation; for a business archive, the company that it is established to serve; the faculty and students for a university. The community that the programme serves imposes its own requirements which directly affects the data selected for preservation, both in their quantity and their nature. For instance, the range will be wider for a national library than for a university-based programme, where it will perhaps be limited to the intellectual output of its faculty and research students.

The community of users will also define the kind of contextual information about data that needs to be preserved. Preserving only the bit-stream is not enough; we must also preserve the additional information and tools needed to access and understand that bit-stream. The UNESCO *Guidelines* provide an example of where documentation needs to be preserved: 'Where digital materials can only be understood by reference to a set of rules such as a record keeping system, database or data generation system, or other contextual information' (UNESCO, 2003, p.76).

Some communities may require only a passive rendition of the data (a screen shot, a PDF version) to be preserved. Other communities of users will need sufficient contextual information to allow the digital materials to be searched or manipulated. Users need 'the option of interrogating old data to produce new results … Some programmes may even have to ensure users can run old simulations, play old computer games, or view digital art in ways that reproduce the original experience rather than a speeded up experience that later technologies may provide' (UNESCO, 2003, pp.77-78). For other communities, enough appropriate contextual information to demonstrate that the authenticity of the data has not been compromised has to be preserved.

The data life-cycle model is here used as a model to examine the role of stakeholders in selection and appraisal decisions. The perspectives of data creation and creators, data creation and creators, and data re-use and re-users are examined.

Data Creation and Data Creators

Increased Role of Stakeholders

Stakeholders will play a greater role in digital preservation than they played in the past in the preservation of non-digital materials. Because selection decisions must be made at an early stage of the lifecycle, an understanding of the needs of the community of users from whom the data originates, who is currently using that data, or who may use it in the future, is needed. **The OAIS** (Open Archival Information System) **reference model** (a widely-adopted standard developed to provide a common framework for describing and comparing architectures and operations of digital archives) **uses the concept of a 'designated community'**. This concept assists appraisal and selection decisions in that what is selected and preserved is independently understandable to the designated community. Data curators need, therefore, to have in mind a group of users, who might change over time. As an example, the traditional distinction between records and archives, in which records become archives only when their active use has ceased and after their value has been ascertained as significant for the future, can no longer apply because data needs active care from its inception. The recordkeeping community has lead in developing responses to digital preservation. Recordkeepers no longer wait 'passively at the end of the life cycle for records to arrive at the archives when their creators no longer wanted them – or were dead' (Cook, 2000, p.2). They need to understand the community for which data is being preserved, and to respond to the needs of that community.

Increasing engagement of some community sectors in the selection of data for

preservation is observable. Greater stakeholder input is being encouraged in many areas. For example, the National Archives of Australia introduced a 'stakeholder consultation approach' to appraisal (Schwirtlich, 2002, pp.60-61). Determining the value of retaining scientific data sets has been determined by peer review (National Research Council, 1995, p.34). Trends such as 'authority by community' (Janes, 2003, p.92), exemplified by wikis, may be predisposing users and other stakeholders to expect greater input into the appraisal and selection process. Members of communities are increasingly playing a role in keeping data accessible. For instance, in universities, faculty members play an expanding role by participating in the establishment of their institutions' e-repositories (Smith, 2003, pp.13-14).

Discipline-based Approaches to Selection

Discipline-based communities are proving effective in determining which data needs to be maintained for use in the future and which additional information has to be integrated in this process. Examples include the Astrophysics Data System (adswww.harvard.edu) and the Australian Bright SPARCS project for the history of science (www.asap.unimelb.edu.au/bsparcs/bsparcs home.htm). For scientific data, there appear to be significant differences in how appraisal criteria need to be applied in different disciplines (Hodge and Frangakis, 2004, p.59), and this is an area worth further investigation.

Technical Capacity to Preserve as a Selection Factor

Brief comment is needed about whether our current technical ability to preserve data should be taken into account when selection and appraisal decisions are made. For paper-based and other traditional materials, as has been noted, a period of non-intervention does not usually result in irreversible damage to or loss of that material: the selection decisions can be separated from considerations of whether we know how to preserve the materials. For data, the two considerations are much more closely intertwined. We know how to maintain data for relatively short periods of time, for example until the data needs to be migrated, or the length of time the storage media remains stable, or the period of time we are willing to maintain software and hardware to read the data. On this basis, then, we may make a decision to select in the awareness that we currently do not yet have the technical processes in place to maintain it for long periods. There will, then, most likely be a need to re-appraise this material when out technical capacity to preserve is more advanced. Until we are more secure in our technical abilities to preserve data over long periods, we need to keep this distinction in mind.

Legal Issues as a Selection Factor

Intellectual property rights and other legal issues assume greater prominence in the preservation of data. Although this applies to all stages in the data life-cycle, it is particularly relevant at the initial (creation, active use) and middle (archiving, preservation) stages. In some data archives selection and acquisition is determined by legal deposit legislation, through which material comes automatically to a designated archive without the expenditure of considerable effort and resources to acquire that material. Most legal deposit

legislation predates the digital era, except in a handful of countries, and changes to such legislation to accommodate data is often considered as an important initial step in a viable national digital preservation programme. Countries that have enacted legislation include Denmark (1998), New Zealand (2003), and the United Kingdom (2003). Other countries have legal deposit legislation that covers some digital materials, typically static publications such as those issued on CD-ROM.

Copyright provisions also have the potential to influence data appraisal and selection decisions. This is noted in the next section.

Data Curation and Data Curators

The Influence of Copyright Legislation on Appraisal and Selection
At later stages (archiving, preservation, access, re-use) in the data life-cycle, **legal issues can also influence appraisal and selection decisions**. For instance, copyright laws usually include a provision that an item can be copied for preservation purposes without specific approval from the copyright owner. Copyright legislation typically does not extend this blanket provision to copying data, so for data covered by such legislation the ability of data curators to make copies for preservation purposes is compromised. This is of crucial importance because copying is the basis of the digital preservation strategies of refreshing, migration, and emulation. Copying data for preservation purposes can infringe current intellectual property rights for some material (Muir, 2004a, pp.76-77). This has been recognised in the copyright legislation of some countries, for example in the United States' Digital Millennium Copyright Act (1998) and in a 1997 amendment to the Canadian Copyright Act, which allow digital materials to be copied if their format has become obsolete (Muir, 2004b, p.72). If copyright and other legal rights are so restrictive that there is no real possibility of access to data being made available in the future, then it is probably pointless to expend resources on its preservation (UNESCO, 2003, p.77).

Metadata's Role in Appraisal and Selection
A digital object consists of much more than just content. It also comprises information that tells us what we need in order to preserve it (to clearly identify it, and to understand the environment in which it was created), information about its attributes (such as file formats), and so on. This information – **metadata** ('structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource' (NISO, 2004, p.1)) **is an integral part of data preservation strategies**. It is, for example, an intrinsic part of the current key digital preservation strategies of emulation, migration, and encapsulation (Day, 2004, p.255).

On the PADI web site preservation metadata is defined as 'structured ways to describe and record information needed to manage the preservation of digital resources'. Preservation metadata stores

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 action' (Preservation metadata, 2003).

In short, preservation metadata is essential for preservation and for possible reuse. Appraisal and selection decisions therefore need to take account of whether or not it is present, and whether sufficient of it, and the right kind, is available. Standardisation of metadata description is one of the most important actions to ensure preservation of data.

Data Re-use and Re-users

Exactly What Are we Preserving?: Essential Elements
Relevant to all stages of the data life-cycle, but particularly germane to the question of re-use of data and therefore included in this section, is the question of exactly what it is we are attempting to preserve. We need to understand the characteristics that data embodies, and **which of these characteristics it is necessary to maintain so that the data can be recreated and re-used in the future**.

The concept of *essential elements* assists. Not every element of a digital object is equally important in recreating it for re-use. Understanding the needs of the user community to whom that data is relevant assists in deciding on which elements are essential. The needs of the community determine the kind of material selected for preservation and the levels of authenticity required. The essential elements (also referred to as *essence* and *significant properties*) of the materials selected are defined in relation to the community's requirements. Some communities place high value on authenticity, so of paramount importance is maintaining the integrity of data: ensuring that any alterations made are carried out only by authorized personnel and are appropriately documented, or that the records are preserved in an unalterable (read-only) form (UNESCO, 2003, p.77). On the other hand, some communities do not require that authenticity is proved to this extent.

Selecting Essential Elements
Questions to be posed in deciding on essential elements to preserve include:

- For whom should this material be kept? Do they have specific expectations about what they will be able to do with the material when it is re-presented?

- Why are the materials worth keeping? What gives them the value that warrants the trouble of preserving them? Is that value associated with evidence, information, artistic or aesthetic factors, significant innovation, historic or cultural association, what a user can make the material do or do with the material, culturally significant characteristics?

- Is the value tied to the way the material looks? (Would it be lost or significantly degraded if the material looked different?)

- Is the value tied to the way the object works? (Would it be lost if particular functions were removed? Or if particular functions happened at a different speed or required different keystrokes?)

- Is the value tied to the context of the material? (Would it be lost if links embedded in the material did not work? Or if a user could no longer see evidence

that connected the material with its original context?)

- Is it possible to distinguish between elements within each of these areas? For example, would advertising banners be considered an essential part of the way the material looked? Would some navigation elements or display functions be needed but not others?

- If it is difficult to define what needs to be maintained, it may be easier to consider the impact of an element not being maintained, and to look for functions or elements that are definitely not needed (UNESCO, 2003, pp.77-78).

Take the example of e-mails. It could be decided that users require only the content information – 'the name and address of the sender, subject, date and time, recipients, and the message, in a standardised structure with only the most simple of formatting' (UNESCO, 2003, p.77). For digital objects harvested from the web, Clausen identifies five aspects of 'preservation quality' – readability, comprehensibility, appearance, functionality, and 'look and feel' – and gives examples of how these might be applied to a range of digital objects commonly encountered on the web (Clausen, 2004, pp.8-10).

Re-appraisal

At the end stages (disposal, transfer of stewardship) of the data life-cycle, re-appraisal of data may be required. Conway notes that 'selection in the digital world is not a choice made once and for all near the end of an item's life cycle, but rather is an ongoing process intimately connected to the active use of the digital files' (Conway, 2000). Re-assessment of the appraisal and selection decisions may be required in order to accommodate changing societal requirements.

A Generic Appraisal and Selection Framework

Given the long list of issues that need to be considered when appraisal and selection is discussed, a generic framework to assist in making selection and appraisal decisions is helpful.

Guidance from the Literature

From the relatively sparse literature on selection of digital materials for preservation, and the considerably larger body of literature about appraisal of electronic records, it is possible to identify some criteria, strategies and typologies as the basis of a generic framework. Much of the literature refers back to a few key sources, such as the CEDARS Project report (2002) and a 2003 CODATA/Erpanet seminar (http://www.erpanet.org/events/2003/lisbon/; Esanu et al, 2004). Much of the literature is about the selection of analogue material for digitisation which, although not the same as selection of data for preservation, provides useful advice for developing appraisal and selection criteria for preservation of data. Gertz, for example, identifies the criteria most frequently cited as:

- Does the item or collection have sufficient value to and demand from a current audience to justify digitization?

- Do we have the legal right to create a digital version?

- Do we have the legal right to disseminate it?

- Can the materials be digitized successfully?
- Do we have the infrastructure to carry out a digital project?

- Does or can digitization add something beyond simply creating a copy?

- Is the cost appropriate? (Gertz, 2000, p.104).

The *context* in which the digital materials are created and used needs to be taken into account in selection and appraisal decisions, as indicated in the questions posed in the Archives Association of Ontario's brief statement about appraisal (Archives Association of Ontario, 2001), which include 'What is the administrative, evidential or informational value of the records to the organization?' and 'Do the records meet the terms of your mandate and acquisition policy?' Also on *context*, the Cedars Project Team report suggests primary selection criteria that include 'currently high use' and 'tied to the long-term or cultural interests of the organisation (Cedars Project Team, 2002, section 4.5.1). Cedars also notes *technical* issues that form part of a selection framework: format issues ('some formats harder to preserve than others'), and technical issues ('technical capacity to preserve may be lacking', 'some technical environments may be easier to preserve than others'); and also suggest the importance of *legal* rights in 'legal status – IP rights need to be negotiated for preservation purposes' (Cedars Project Team, 2002, pp.109-110). The *Decision Tree for Selection of Digital Materials for Long-term Retention*, in the influential Digital Preservation Coalition's handbook on the preservation of digital materials (Jones and Beagrie, 2001, section 4) poses questions in four categories:

1) selection *policy*: is there an institutional selection policy? Does the material fit into it? Is the material of long-term value?

2) *legal and intellectual property issues*: have acceptable rights been negotiated? Can they be?

3) *Technical* questions: can you handle the file format, now and in the future? Can the material be transferred to a more manageable format?

4) The existence of *documentation and metadata*: has sufficient been supplied?

Risk Management

Discussions about appraisal and selection are increasingly being based on the concept of risk management. Given that it is unrealistic to expect that sufficient resources will be available for preserving all of the data identified as significant, the risk management approach assists. It allows us to refine decision-making by balancing the risks of reduced accessibility to or loss of materials against the consequences of that reduced accessibility or loss. If material falls into the category of 'high risk and high consequence', it becomes the prime candidate for preservation attention. If, on the other hand, it is categorized as 'low risk, low consequence', it receives the least attention. The PADI web site provides further information about risk management (Risk management 2005).

The Generic Framework

Table 1 (in Section 11) is based on the literature referred to in the section immediately above, as well as other

literature. It tabulates selection criteria applied to traditional (non-digital) material, expands these from the literature about selection for digitising, then notes criteria from two key documents, the CEDARS report (2002) and the DPC Decision Tree (Jones and Beagrie, 2001, section 4). Table 1 indicates that changes are required to the appraisal and selection procedures developed for non-digital artefacts to accommodate data. While some factors assume greater importance, new factors must be added. The digital selection frameworks developed to date still place high priority on criteria for determining value, but also strongly emphasise other criteria such as the legal and intellectual property rights governing a resource, whether we have the technical ability to preserve it, the costs involved in preserving it, and the presence of appropriate documentation and metadata.

We need to expand upon this starting point. Increasingly non-traditional criteria, such as the legal and intellectual property rights governing a digital object, whether we have the technical ability to preserve it, the costs involved in preserving it, and the presence of appropriate documentation and metadata, are becoming central to appraisal and selection decision-making. Over-riding all of these, however, is the need to closely link appraisal and selection criteria to particular communities of data users. The next section notes some current examples of appraisal and selection criteria.

Appraisal and Selection in Action

Scarcity of Data Appraisal and Selection Policies

Despite the literature's strong emphasis on the need for appraisal and selection policies, surprisingly few of these have been articulated and made public. Electronic recordkeeping documentation makes interesting reading in this respect: it typically refers to appraisal assumptions, but they are not stated in the documents, and one is lead to wonder whether they are clearly articulated elsewhere. Web documents about digital preservation often simply refer back to existing non-digital library collection criteria, but, as noted above, these are insufficient for data preservation. This section examines some of the policies that are available, in three areas: libraries and digital libraries, recordkeeping, and research data archiving.

Selection Policies and Criteria: Libraries and Digital Libraries

Policies that refer to selection criteria for long-term preservation are not widely available publicly. This section refers to some that have been developed for selection of digital materials for long-term retention in libraries and in digital libraries.

SunSITE Digital Library

One of the earliest policies located is that for the University of California Berkeley Library's SunSITE Digital Library, dated 1996 (University of California Berkeley 1996). 'Archival status' is assigned to material that has been considered against these criteria:

- Perceived usefulness of the material

- Perceived life-span of the material (is it likely to always be of significance?)

- Availability of the material elsewhere (Is the SunSITE the only location that is hosting it?)

- Uniqueness of the material (Are we the only institution that **can** host the material? e.g., primary sources)

- Commitment by another institution to archive the material.

Library of Congress

The Library of Congress provides a fuller statement (Library of Congress (1999). *Collections Policy Statements.* http://www.lcweb.loc.gov/acq/devpol/electron.html). Its criteria for selecting 'electronic resources' for its permanent collection 'do not greatly differ from those used for books or materials in other formats', in that 'the cost of the work and the requirements of serving, cataloging, storing, and preserving must be considered in the decision'. General criteria are specified: usefulness in serving the current or future informational needs of Congress and researchers, reputation of the information provider, amount of unique information provided, scholarly content, currency of the information, frequency of updating, and ease of access. In addition, specific guidelines are considered:

- **Content.** Give priority to items which will be of the greatest current or future use to Congress and or to the greatest number of researchers and staff.

- **Added Value.** If the material in electronic format is also available in print, the electronic resources should provide added value over its print equivalents, including timely access, lower costs, enhanced searching, or access from multiple workstations. The ability to make the resource available on a network among reading rooms in different buildings is a high priority.

- **Ease of Use.** The work should be easy to use, requiring minimum training. Documentation supplied by the vendor must be clear. Expensive items should be evaluated with an on-site pilot.

- **Maintenance.** The amount of support required by staff to make the resources available must be considered. The decision to collect resources requiring significant amounts of staff time to preserve, including migration to newer formats, must be weighed against the current and future scholarly value of the resources.

- **Standards.** The work should meet acceptable, commonly used technical standards, digital formats, and practices.

- **Equipment.** The work should operate on equipment and operating systems either currently or expected-to-be available. Resources requiring extensive, specialized, and/or expensive new equipment or storage space to make them available will be acquired only if the research value is indisputably high.

- **Output.** The work should provide convenient output to printers and/or users' files.

University Libraries

By comparison, the Columbia University Libraries' *Policy for Preservation of Digital Resources* (Columbia University Libraries (2000) does little more that state

that selection criteria for digital resources are the same as for all CUL collections, and then restates those very broad criteria later in the policy. This is not uncommon: another example is Cornell University Library's Digital Preservation Policy Framework which lists some selection criteria (in Appendix C) which are essentially the same as for all collections and refers to other policies (Cornell University Library (2004).

Genre Selection Policies

A more specific statement is *How to Identify the "Best" Resources for the Reviewed Collection of the Digital Library for Earth System Education* (Kastens 2002?). This specifies seven selection criteria: resources must be

- Scientifically accurate
- Important or significant
- Pedagogically effective: has student learning occurred?
- Well-documented
- East of use for students and faculty
- Able to inspire or motivate students
- Robust and sustainable as a digital resource.

This set of criteria is of particular interest for two reasons. First, it clearly links a particular user community and their specific requirements. Second, it indicates in more detail the metrics that are used for these criteria, by coordinating standard measures used in other contexts:

Scientific accuracy is most commonly assessed by peer review by scientist-experts recruited by a journal editor. Pedagogical effectiveness is commonly evaluated by pedagogy-experts through classroom observation, interviews, questionnaires, and other instruments. Robustness of a digital resource is commonly evaluated by QA (quality assurance) specialists as part of a software development effort.

More selection criteria for specific genres of data, especially web sites, are available. Selection criteria for the National Library of Australia's PANDORA archive of Australian online publications are thoroughly articulated, as are that Library's policies for other genres (National Library of Australia 2001?):

- for the PANDORA archive of Australian online publications, selection decisions are guided by detailed selection guidelines ...

- for Australian physical format digital publications, preservation decisions are guided by detailed guidelines …

- for unpublished computer files deposited in the Library's Manuscripts collection, selection decisions are made on a case by case basis, depending on what is known about the material, its expected uniqueness and significance, and the technical difficulties of transferring or retrieving data ...

- for digital spatial data sets and mapping resources acquired for the Maps collection, a basis for selection decisions is still being developed.

- for the Library's corporate records in digital form, selection decisions are guided by a records disposal schedule approved by the National Archives of Australia.

- for information resources listed on the PADI subject gateway, 'Safekeeping'

selection decisions are guided by detailed guidelines.

- for metadata records of information resources that are required for long-term retention, all items are selected for preservation.

Other selection documents are available, such as the British Library's *Collection Development Policy for UK websites* (British Library 2004) which includes selection criteria, and Nicholls and Williams (2002) which lists criteria for selecting web pages for retention at the University of Melbourne, Australia:

1. Is the content of the webpage published in another format?
    - If answer is YES – check to see if the content between the two is significantly different. If the content is the same, then retain only the paper copy. If the content is different then check against the criteria listed below.
    - If the answer is NO – check it against the criteria listed below.

Does the webpage:

2. Publish a change in policy?
3. Create a publishing precedent for the University (i.e. is this the first time the material on the site has been published on the web)?
4. Represent a substantive business of the work unit, section or University?
5. Publish legal advice?
6. Publish information involving negotiations on behalf of the University?
7. Does the webpage transmit formal communication(s) between officers?
8. Has the webpage been used to initiate, continue or complete a departmental activity/transaction?
9. Does the webpage have continuing value for others in the work unit?
10. Does anyone external to the work unit need to be aware of, or refer to, this webpage for evidentiary purposes now or in the future?

If the answer is YES to any of the above, retain for long-term preservation.

E-repositories

E-repositories are an interesting case in relation to selection. They usually involve self-selection and self-contribution by authors, but institutional management (as one example see Cornell University's repository (http://dlist.sir.arizona.edu/903/01/CUL%2BInstitutional%2BRepositories%2Bfor%2BALA.pdf). E-repositories provide examples of selection for a specific user community with significant input from members of that community: as noted above, trends such as 'authority by community' and wikis may be predisposing users and other stakeholders to expect greater input into the appraisal and selection process.

These examples are only a small selection from a larger number available publicly (more can be viewed on the PADI web site (Selection 2005)) and there are without question many more available only as in-house documents. However, it is unlikely that they are significantly more developed than those noted above.

Appraisal Policies and Criteria: Recordkeeping

Recordkeeping Principles

Recordkeeping has a long tradition of appraisal theory and practice. This tradition has been investigated for its applicability to electronic records, most notably in several research projects under the InterPARES umbrella which has developed a statement of *Benchmark Requirements for Assessing the Authenticity of Electronic Records* provide (InterPARES Authenticity Task Force 2002). Outcomes such as these have lead to new concepts of appraisal. This 'new' appraisal can be seen, for data, to apply at a number of 'critical appraisal points':

- capture – determining what records to vest with additional protection of recordkeeping processes to carry them through time;

- reach – determining how far the records should be intelligible outside their immediate domain of capture (and this process keeps on going in an iterative fashion, to enable records to be understood over time, as well as over physical/virtual space);

- migration – determining what records should be maintained in a usable format over time and through system changes; and

- destruction/retention – determining which records to retain and for how long, itself a multiple decision-making process inside whatever structure is relevant, be it a small group, a larger department, a whole organization or the more traditional archival approach involving retention beyond organizational boundaries (Reed 2005, p.125).

As noted above, all appraisal decisions are necessarily subjective; re-appraisal is a consequence of this inevitability.

The literature about the relevance of archival principles to data curation is well worth reading for its perspectives. Two helpful pieces are Gilliland-Swetland (2000) and Menne-Haritz (1999).

Non-specific Nature of Appraisal Policies

Most publicly available appraisal policies for electronic records assume the application, at an early stage in the record life cycle, of standard archival retention and disposal schedules that are the outcome of appraisal, and are not specific about criteria for long-term retention of data. For instance, Edinburgh University Archives' undated *Archival Selection Criteria* makes no mention of archives and records format, and one can only assume that it is intended to cover electronic records. The Committee on Institutional Cooperation's University Archivists Group's *Standards for an Electronic Records Policy* (2001) simply indicates that 'Records, including electronic records, shall be retained or disposed of in accordance with authorized and approved records retention schedules'.

Macro-appraisal

The macro-appraisal concept is another outcome of the recordkeeping community's attempts to apply appraisal concepts to data. This attempts to refine decisions made about value. It

is nothing more than deciding (what records to create and) how long they should be kept or deciding what archival

records to collect, *by first* mapping the territory; *by first* identifying and analyzing the theoretical documentary universe … Only *after* that intensive research based on examination of societal or business functions … should you do 'micro-appraisal' by identifying the key records (Piggott, 2001).

This approach, it is suggested, helps cope with greater quantities of records. It is also being applied by involving user communities more closely in appraisal decisions. Statements about macro-appraisal include those by the National Archives of Australia (National Archives of Australia 2005, Harris 2005) and the U.K. National Archives (National Archives (UK) 2004, which includes an Appendix of 'Appraisal values' and another noting 'Summary of change in timing of application of appraisal criteria', and Mercer 2004).

Appraisal and Selection Policies and Criteria: Research Data

Research Data: A Special Case?

It is frequently argued that all research data can be considered as an important national resource and should be retained indefinitely because of its potential to be re-analysed using different parameters or new techniques. (An example is the chapter about data management in the US Climate Change Science Program's *Strategic Plan: Final Report* (2003) which includes the comment that 'new technologies need to be developed that will enable us to keep *all* data needed for long-term global change research, reducing the need to prioritize which data will be archived.') A major theme of the literature about data management in science archives is open access to science data, which includes some coverage of preservation issues and reference to the probability that it won't be possible or useful to keep everything. However, beyond this recognition little of specific assistance in appraising and selecting research data is publicly available. Relatively few commentators suggests that not all data should be kept. Some that do are Beedham and others in 'The Selection, Appraisal and Retention of Digital Social Science Data' (Beedham *et al* 2004) and the JISC *e-Science Curation Report* (Lord & Macdonald 2003).

Examples of Selection Criteria for Research Data

Some statements about appraisal and selection criteria for research data are publicly available. The National Environment Research Council's *Data Policy Handbook* (2002) contains much that is relevant about ensuring that the right data sets are kept and about who has the responsibility for identifying what those are, although it does not provide a list of criteria. Albeit in a specific narrow social science research data area, criteria for inclusion in the *Sociometrics Corporation* Research Archive on Disability in the U.S. provide guidance (Sociometrics 1999); these are a mixture of technical and content criteria. The International Federation of Data Organizations for the Social Sciences has attempted to list criteria for selecting datasets for retention (Mochmann 2005). Possible retention criteria for epidemiological data sets are suggested in the JISC *E-science curation Report*:

The nature of the questions being asked by the study

- Whether it addresses only one question or many

- Whether the question has been asked before
- The richness of the data set
- If it is a longitudinal study – 'indicates an amber light'
- Sample-related studies
- Stability of the measures used
- Possibility to go back to the population (e.g. for consent, ethical committee access)
- Uniqueness; value for possible future comparisons (Lord & Macdonald 2003, p.46).

Research datasets may be an excellent case where re-appraisal at defined intervals is particularly applicable. Most could be initially kept, then re-appraisal occurs at defined intervals to test the dataset against agreed-upon criteria to establish whether it still meets the conditions for applying resources to its long-term retention.

The State Library of Victoria, Melbourne, Australia

The State Library of Victoria has articulated its appraisal and selection criteria for electronic resource. To support its policy of providing access to digital materials in preference to collecting 'other forms containing like information', it has developed *Digital Library Collection Development Guidelines* (2005) to supplement its development policies for print and analogue material (State Library of Victoria 2001). The digital guidelines first indicate six categories of material the Library collects: purchased or licensed material (such as electronic journals or databases), deposit material (such as government publications), 'links and pointers to free Internet resources where the URL is added to the Library's catalogue, material digitised by the Library from its collections, material digitised by other organisations and acquired by the library, and online Victorian publications that have been selected for preservation in the national cooperative PANDORA scheme (see section 5.2.4). The selection guidelines that are applied have four areas: **content, format appropriate to the content, practical issues, and strategic considerations**. *Content* is determined in relation to the Library's collection development policy. *Format* is determined by considering what is most suitable for the Library's purpose (for example, reference material is preferred in a format able to be networked throughout the Library). *Practical issues* encompass budgetary, technical and legal constraints, including stability of the format. *Strategic considerations* relate to the Library's overall directions and information technology plans.

The State Library of Victoria has also developed other policies to assist with selection of digital objects for preservation, such as a *Digital Collection Preservation Plan* and a draft policy for *Digital Preservation Procedure* (both unpublished). The latter articulates a detailed assessment process for determining preservation priorities for digital objects. These priorities could change over time as a result of re-appraisal. At the acquisition stage these materials are categorised as 1 (high priority for preservation), 2 (medium priority) and 3 (low priority). Three questions are posed:

1. Is the Digital Object significant? Yes? Score 1 point

2. Is the Digital Object vulnerable? Yes?

Score 1 point

3. Is the Digital Object Scarce? Yes? Score 1 point

Three points equates to high priority, two points to medium, and one or zero points to low priority. Significance, vulnerability and scarcity are determined by considering further questions. A digital object is **significant** if the answer to one or more of the following questions is yes:

a. Does it have Victorian content?
b. Does it have Australian or international content of special interest to Victorians?
c. Is it a 'standalone' legal deposit item? (that is, it does not come with a book or other publication).

A digital object is **vulnerable** if the answer to one or more of the following questions is yes:

a. Does it require special hardware to load?
b. Does it require special software to load?
c. Is the media more than 15 years old?
d. Is the object on an obsolete media (e.g. 5.25 inch floppy)?
e. Will it be in an area where it is at risk of theft or damage?

A digital object is **scarce** if the answer to one or more of the following questions is yes:

a. Is it unique?
b. Is there something very peculiar or original about the type of object?
c. Is it almost impossible to obtain another copy?
d. Is it an expensive (over $500) publication?

Next steps

Developing an Institution-specific Selection Framework

Guidance to assist in developing a selection framework is found in the questions posed in the DPC Handbook's *Decision Tree for Selection of Digital Materials for Long-term Retention*):

- *Policy*: does the institution have a selection policy? Does the material fit into it?
- *Legal and intellectual property issues*: have, and can, acceptable rights be negotiated?
- *Technical* issues: can the file format be handled, currently and in future? Is transfer to a more manageable format?
- The existence of *documentation and metadata*: is there sufficient? (Jones & Beagrie 2001: section 4).

The appraisal toolkit for electronic records of the Public Records Office in the United Kingdom (Public Record Office, 2000) also provides guidance.

Key Questions and Actions

This section provides a list of key actions and questions to assist in developing and implementing an institutional appraisal and selection policy for data preservation.

1. Investigate your environment. Ask questions such as:

- *What is the context?* What is the institution's mission? Goals? Resource

limitations – people with knowledge and time, facilities, equipment to examine material?

- *What are your technical capabilities?* Is the technical knowledge and equipment to maintain data once it is ingested or acquired available?

- *What are your legal obligations and rights in relation to your data?* Are there legal requirements for retaining data for specific periods of time? Do other legal requirements (such as intellectual property rights) restrict the data that can be preserved?

2. Develop an appraisal and selection policy. A policy allows informed, consistent and accountable decisions about appraisal and selection to be made in situations where judgments are subjective and speculative (UNESCO 2003,12.7). This policy should include a statement of re-appraisal principles and a re-appraisal schedule.

3. Develop specific criteria for your context about what data to keep, which elements of the data are essential, and what documentation is required. Key questions include:

- *What data to keep?* 'Decisions should be based primarily on the value of material in supporting the mission of the organisation taking preservation responsibility' UNESCO 2003, 12.8), but many other factors (noted in above sections) must also be taken into account, such as the cost of preservation, and its technical difficulties.

- *What elements of the data to keep?* Knowing the user community's requirements is essential for making informed decisions about what the essential elements of data are likely to be. The UNESCO *Guidelines* suggest the following questions:

- For whom should this material be kept? Do they have specific expectations about what they will be able to do with the material when it is re-presented?

- Why are the materials worth keeping? What gives them the value that warrants the trouble of preserving them? Is that value associated with:

  – Evidence
  – Information
  – Artistic or aesthetic factors
  – Significant innovation
  – Historic or cultural association
  – What a user can make the material do, or do with the material
  – Culturally significant characteristics?

- Is the value tied to the way the material looks? (Would it be lost or significantly degraded if the material looked different?)

- Is the value tied to the way the object works? (Would it be lost if particular functions were removed? Or if particular functions happened at a different speed or required different keystrokes?)

- Is the value tied to the context of the material? (Would it be lost if links embedded in the material did not work? Or if a user could no longer see evidence that connected the material with its original context?)

- Is it possible to distinguish between elements within each of these areas? For example, would advertising banners be considered an essential part of the way the material looked? Would some navigation elements or display functions be needed but not others? (UNESCO 2003, 12.20).

- *How much documentation to keep?* Selection processes should identify the documentation that will be needed to make data understandable in the future: for example, software manuals may be required to understand how data should be presented.

4. Engage your stakeholders. An example is encouraging members of a user community to assist in selection decision-making, for instance to determine value. Another example is the role of data creators in selecting what should be preserved and in providing metadata at the point of creation.

Content, Structure and Context

Perhaps the most crucial step in developing workable appraisal and selection guidelines for data preservation is to identify the essential elements of data in relation to its specific user community. The concepts of content, structure, and context may assist. (This idea is further explained in Harvey 2005a and Harvey 2005b, chapter 4). Context in particular is increasingly being recognised as essential to document and maintain: 'Digital scientific data depend on the perpetuation of context to ensure their long-term value and usability. Without it, the data are essentially meaningless' (ERPANET/CODATA 2003.)

The example of videotapes maintained by a government-owned broadcasting company's library, although not digital material, illustrates these concepts. Material of national heritage significance is produced, so *content* is significant. The *structure* (record form) becomes a crucial criterion for selection., because technological obsolescence of recording and playback equipment for various videotape formats is a major issue. Selection criteria should consider the 'obsolescence rating' for each format, such as *lower risk* for formats such as VHS, and *critically endangered* for 1-inch SMPTE (Ampex). There is a legal deposit obligation for some material produced, so the *context* in which these videotapes were created must be considered. Another example is dissertations in digital form. Here the *structure* (record form) is created by word-processing and/or imaging software, for which documentation needs to be retained. *Contextual* information to retain includes documentation that explains the reasons why and the conditions under which this material was created. These examples are summarized in Table 2 (see section 11).

Future developments

New uses are being made of digital materials, in particular the high value being placed on the ability to reuse and repurpose data. These are altering the way in which value and significance are perceived, and are significantly affecting how data is appraised and selected for preservation. Changing, too, are institutional structures and modes of information production, and these factors, plus others, are rapidly modifying data curation practice in ways that are as yet not fully clear. Despite such uncertainties, four areas can be identified that will

contribute to improvements in appraisal and selection practice in data curation: further research, refinements of current appraisal and selection theory and practice, development of best practice guidelines, and greater engagement of stakeholders.

Research into appraisal and selection is needed, as it is into other parts of the data life cycle. Such investigation need to take into account the high amounts of human input required in current processes, and their cost. Identifying selection processes that can be automated may make it possible to reduce the level of resources needed for selection. Other appraisal and selection issues where research is likely to result in new advances include: better defining how digital materials are used; defining the essential elements for categories of data; establishing how, and when, re-appraisal should be carried out; identifying staged selection and appraisal procedures; and how to make selection and appraisal routine activities in developing and managing digital libraries. The outcomes may well result in 'radically different approaches' (Hedstrom et al, 2003, p.7; see also Ross and Hedstrom, 2005).

Further refinements of appraisal and selection theory and practice, such as macro-appraisal and other 'new' appraisal concepts (see section 5.3) are likely to assist in improving practice. The lead will come from the recordkeeping sector, through research projects such as InterPARES.

The development of best practice appraisal and selection guidelines for a range of data curation sites, promulgated widely to practitioners, will assist in improving practice. Allied to this is the need for greater engagement of stakeholders in developing guidelines for selection and appraisal and also in making selection decisions, as noted above in sections 4.2 and 4.3.

Conclusion

A commentator on an ERPANET workshop about digital preservation business models held in 2004 noted several key themes relating to selection and appraisal:

- The renewed importance of selection/appraisal
- The need to make value judgements about what is important enough to be preserved so that effective resource allocation can be made
- The strong link between funding for preservation and requirements for access: 'Sustainability is more likely to be achieved if we tie our preservation strategies to access strategies that meet the short-term information needs such as legislative requirements, risk management, and the potential to exploit digital assets commercially'
- The concept of 'resource scarcity': it is essential to recognize that there won't be sufficient resources to preserve all (Searle 2005).

While it is reasonably straightforward to identify the criteria traditionally used by libraries when they select for preservation, and to describe appraisal practice for archival material, these cannot be applied directly to the selection of digital materials for preservation. What is needed is a combination of existing criteria, weighted

differently, to which new criteria are added. As one example, Eastwood suggests that the cost of preservation of digital materials will become 'rather more determinative of the outcome of appraisal' than it was for traditional materials (Eastwood, 2003). The existing criteria will bear closer scrutiny to glean from them principles and practices that can usefully be applied to data preservation; here, the leads demonstrated by applying recordkeeping concepts to data preservation are promising. There is now sufficient experience with data curation to begin to develop practical guidelines, which will be based on statements such as the ones described in this chapter. Improved appraisal and selection practice will develop as data curators become increasingly aware of the practice of others and of the importance of making informed and defensible selection decisions.

The challenges of appraisal and selection are considerable, in part because preservation is 'a relative rather than an absolute concept because objects change over time as do our approaches to viewing or interpreting those objects' (Cloonan, 2001, p.235). We are not likely to get it completely right, but it is, nevertheless, our professional responsibility to rise to the challenges posed with some understanding of the consequences of ignoring them.

Terminology

Appraisal: 'the process of evaluating records to determine which are to be retained as archives, which are to be kept for specified periods and which are to be destroyed' (Ellis, 1993, p.461).

Authenticity: 'Quality of genuineness and trustworthiness of some digital materials, as being what they purport to be, either as an original object or as a reliable copy derived by fully documented processes from an original' (UNESCO 2003)

Essential elements: 'The elements, characteristics and attributes of a given digital object that must be preserved in order to re-present its essential meaning or purpose. Also called *significant properties* by some researchers' (UNESCO 2003)

Metadata: 'structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource' (NISO, 2004, p.1)

Preservation metadata: 'Metadata intended to support preservation management of digital materials, by documenting their identity, technical characteristics, means of access, responsibility, history, context, history and preservation objectives' (UNESCO 2003)

Records: in the recordkeeping context, records are evidence of transactions

Risk management: 'Process of identifying and assessing risks presented by threats, and if appropriate, taking steps to bring the level of

risk down to an acceptable level' (UNESCO 2003)

Selection: the process of deciding what items or resources will be added to a library's collection.

References

Archives Association of Ontario (2001) What to keep and what to destroy? aao.fis.utoronto.ca/aa/appraisal.html

Beedham, H. et al (2004) The Selection, Appraisal and Retention of Digital Social Science Data, *Data Science Journal* 3: 209-221.

British Library (2004). *Collection Development Policy for UK websites* http://www.bl.uk/collections/britirish/modbritcdpwebsites.doc

Burrows, T. (2000) Preserving the past, conceptualising the future: research libraries and digital preservation. *Australian Academic & Research Libraries*, **31** (4), 142-153

Cedars Project Team (2002) The Cedars Project Report, April 1998-March 2001 www.leeds.ac.uk/cedars/OurPublications/cedarsrepmar01exec.html

Clausen, L.R. (2004) *Handling file formats*. Århus: State and University Library

Cloonan, M.V. (2001) W(h)ither preservation? *Library Quarterly,* **71** (2), 231-242

Columbia University Libraries (2000). *Policy for Preservation of Digital Resources.* http://www.columbia.edu/cu/lweb/img/assets/4776/dlpolicy.pdf

Committee on Institutional Cooperation, University Archivists Group (2001). *Standards for an Electronic Records Policy* http://www-personal.umich.edu/~deromedi/CIC/cic4.htm

Conway, P. (2000) Overview: rationale for digitization and preservation. In *Handbook for digital projects: a management tool for preservation and access*, ed. M. Sitts. Andover, Mass.: Northeast Document Conservation Center www.nedcc.org/digital/dighome.htm

Cook, T. (2000) Beyond the screen: the records continuum and archival cultural heritage. Paper presented at the *Australian Society of Archivist Conference,* Melbourne, 18 August 2000. www.archivists.org.u/sem/conf2000/terrycook.pdf

Cornell University Library (2004). *Digital Preservation Policy Framework.* http://commondepository.library.cornell.edu/cul-dp-framework.pdf

Day, M. (2004) Preservation metadata. In *Metadata Applications and Management* ed. G.E. Gorman and D.G. Dorner, pp.253-273. London: Facet

Edinburgh University Archives (undated). *Archival Selection Criteria.* http://www.lib.ed.ac.uk/resources/collections/specdivision/criteria.pdf

Ellis, J. (1993) (ed.) *Keeping archives* 2nd edn. Melbourne: Thorpe in association with the Australian Society of Archivists

ERPANET/CODATA Workshop (2003) The selection, appraisal and retention of digital scientific data: final report, ERPANET/CODATA Workshop, Biblioteca Nacional, Lisbon, December 15-17, 2003. www.erpanet.org/events/2003/lisbon/LisbonReportFinal.pdf

Esanu, J. et al, (2004). 'Selection, apprailsa, and retention of digital scientific data: highlights of an ERPANET/CODATA Workshop', *Data Science Journal*, **3, 30 December,** 226-232.

Gertz, J. (2000) Selection for preservation in the digital age. *Library Resources & Technical Services*, **44** (2), 97-104

Gilliland-Swetland, A.J. (2000) *Enduring paradigm: the value of the archival perspective in the digital environment*. Washington, D.C.: Council on Library and Information Resources.

Harris, E. (2005) *Macro-Appraisal at the National Archives.* National Archives of Australia. http://www.naa.gov.au/recordkeeping/rkpubs/fora/05May/macro-appraisal.pdf

Harvey, R. (2005a) Preserving digital documentary heritage in libraries: what do we select? In *Preservation of Electronic Records: New Knowledge and Decision-making: Postprints of a Conference, Symposium 203, Ottawa, Canada, September 15-18, 2003, Ottawa, Canada, 15-18 September 2003* (Ottawa: Canadian Conservation Institute), pp. 13-20.

Harvey, R. (2005b) *Preserving Digital Materials*. Munich, K.G. Saur

Hedstrom, M. et al. (2003) *Invest to Save: Report and Recommendation of the NSF-DELOS Working Group on Digital Archiving and Preservation* (Pisa & Washington DC). http://delos-noe.iei.pi.cnr.it/activities/internationalforum/Joint-WGs/digitalarchiving/Digitalarchiving.pdf

Hodge, G. and Frangakis, E. (2004) Digital preservation and permanent access to scientific information: the state of the practice: a report sponsored by the International Council for Scientific and Technical Information (ICSTI) and CENDI. cendi.dtic.mil/publications/04-3dig_preserv.html

InterPARES Authenticity Task Force (2002) *Requirements for Assessing and Maintaining the Authenticity of Electronic Records*. http://www.interpares.org/book/interpares_book_k_app02.pdf

Janes, J. (2003) Internet librarian: authority by community. *American Libraries*, **January,** 92

Jones, M. and Beagrie, N. (2001) *Preservation management of digital materials: a handbook.* London: British Library.

Kastens, K. (2002?) *How to Identify the "Best" Resources for the Reviewed Collection of the Digital Library for Earth System Education.* http://www.ldeo.columbia.edu/edu/DLESE/collections/CGms.html

Library of Congress (1999). *Collections Policy Statements.* http://www.lcweb.loc.gov/acq/devpol/electron.html).

Lord, P. and Macdonald, A. (2003) *E-science curation report: data curation for e-science in the UK: an audit to establish requirements for future curation and provision.* Prepared for the JISC Committee for the Support of Research (JCSR). London: Digital Archival Consultancy

Menne-Haritz, A. (1999) Appraisal and disposal of electronic records and the principle of provenance: Appraisal for access – not for oblivion. http://www.narc.fi/dlm/9b.htm

Mochmann, E. (2005) The social science data archive step by step. http://www.ifdo.org/data/data_archive_workflow03.html

Muir, A. (2004a) Digital preservation: awareness, responsibility and rights issues. *Journal of Information Science*, **30** (1), 73-92.

Muir, A. (2004b) Issues in the long-term management of digital materials. In *Managing preservation for libraries and archives: current practice and future developments*, ed. J. Feather, pp.67-81. Aldershot, Hants: Ashgate

National Archives (UK) (2004) *Appraisal Policy*. http://www.nationalarchives.gov.uk/recordsmanagement/selection/pdf/appraisal_policy.pdf

National Archives of Australia (2005) *What is Macro-appraisal?* http://www.naa.gov.au/recordkeeping/disposal/appraisal/macro-appraisal.html

National Environment Research Council (2002) *NERC Data Policy Handbook.* Version 2.2. http://www.nerc.ac.uk/data/documents/datahandbook.pdf

National Library of Australia (2001?) *Selecting Digital Materials to be Preserved* http://www.nla.gov.au/preserve/digipres/selecting.html

Nicholls, C. and Williams, J-P. (2002) 'Identifying Roadkill on the Information Superhighway: A Website Appraisal Case Study' *Archives and Manuscripts* 30(2), pp. 96-111.

NISO (2004) *Understanding metadata*. Bethesda, MD: National Information Standards Organization Press

Piggott, M. (2001) Appraisal: the state of the art: paper delivered at a professional development workshop presented by ASA South Australia Branch 26 March 2001 [www.archivists.org.au/sem/misc/piggott.html]

Preservation metadata (2003) [PADI summary]. National Library of Australia [www.nla.gov.au/padi/topics/32.html]

Public Record Office (2000) *Evaluating information assets: appraising the inventory of electronic records*. Kew, Public Record Office.
www.nationalarchives.gov.uk/electronicrecords/advice/pdf/appraisal_toolkit.pdf

Reed, B (2005) Records. In *Archives: Recordkeeping in Society*. Wagga Wagga, NSW: Centre for Information Studies, pp.101-130.

Ross, S. and Hedstrom, M. (2005) 'Preservation research and sustainable digital libraries', *International Journal on Digital Libraries*, 5 (4), 317-325.

Risk management (2005) [PADI summary]. National Library of Australia [http://www.nla.gov.au/padi/topics/272.html]

Schwirtlich, A. (2002) The functional approach to appraisal: the experience of the National Archives of Australia. *Comma*, **1-2**, 57-62.

Searle, S. (2005) 'Conference Report: Business Models Related to Digital Preservation', *New Zealand Libraries,* **49**,: 418- 422

Selection (2005) [PADI summary]. National Library of Australia [www.nla.gov.au/padi/topics/9.html]

Smith, A. (2003) *New-model scholarship: how will it survive?* Washington, D.C.: Council on Library and Information Resources

Sociometrics Corporation (1999). Research Archive on Disability in the U.S.
http://www.socio.com/data_arc/radius_0.htm

State Library of Victoria (2001) *Collection Development Policy*. Melbourne: State Library of Victoria.

State Library of Victoria (2005) *Digital Library Collection Development Guidelines*.
http://www.slv.vic.gov.au/about/information/policies/cdp/digital_guide.html

Tibbo, H.R. (2003) On the nature and importance of archiving in the digital age. *Advances in Computers*, **57**, 1-67

UNESCO (2003) *Guidelines for the preservation of digital heritage*, prepared by the National Library of Australia. Paris: UNESCO.
unesdoc.unesco.org/images/0013/001300/130071e.pdf

University of California Berkeley SunSITE Digital Library (1996). *Archived Collection Level.*
http://sunsite.berkeley.edu/Admin/archived.html).

US Climate Change Science Program (2003) *Strategic Plan: Final Report.*
http://www.climatescience.gov/Library/stratplan2003/final/ccspstratplan2003-chap13.htm

Vogt-O'Connor, D. (2000) Selection of materials for scanning. In *Handbook for digital projects: a management tool for preservation and access,* ed. M. Sitts. Andover, Mass.: Northeast Document Conservation Center [www.nedcc.org/digital/dighome.htm]

Appendices

	Traditional Selection Criteria	**Criteria Applied to Selection for Digitizing**	**CEDARS/DPC *Decision Tree***
Value	Evidential Aesthetic Market Associational Exhibition Informational	Sufficient value to current audience Does digitization add value?	Significant long-term value?
Physical condition	Threat to object Fragility		
Resources available	Management plan?	Is infrastructure available? Is cost appropriate?	
Use	Heavy use	Current demand	Currently high use
Social significance	Held in community esteem?		Tied to long-term interests of organization
Legal rights	Copyright	Rights to digitize Rights to disseminate	Legal status IP rights
Format issues		Can it be digitized successfully?	Type of material (can it be digitized successfully?)
Technical issues			Technical ability to preserve? Can file format be handled?
Policies			Selection policy?
Documentation			Sufficient available?

Table 1: Generic Appraisal and Selection Framework
(From Harvey 2005b)

	Structure (record form)	**Content (information)**	**Context (linkages)**
Videotapes	VHS, Digital Betacam (Sony), 1-inch Umatic, 1-inch SMPTE (Ampex)	Significant national heritage content (e.g. documentaries)	Legal deposit regulations; intellectual property rights of other parties
Digital theses	Word-processing software, pdf	Significant intellectual content	Regulations under which submitted; value as a record of the university's activities

Table 2: Content, Structure and Content of Some Digital Materials
(From Harvey 2005b)

Annotated List of Key External Resources

Jones, M. and Beagrie, N. (2001) *Preservation management of digital materials: a handbook. Decision Tree for Selection of Digital Materials for Long-term Retention*)
http://www.dpconline.org/graphics/handbook/figure4.html
Useful flow-chart approach to selection decision making for data preservation.

Selection (2005) [PADI summary]. National Library of Australia [www.nla.gov.au/padi/topics/9.html]
The National Library of Australia's PADI web site gathers together a considerable number of sources about appraisal and selection for data preservation.

UNESCO (2003) *Guidelines for the preservation of digital heritage*, prepared by the National Library of Australia. Paris: UNESCO.
unesdoc.unesco.org/images/0013/001300/130071e.pdf
Chapter 12 "Deciding What to Keep") is essential reading.

Cornell University Library (2003) *Digital Preservation Management: Implementing Short-term Strategies for Long-term problems* http://www.library.cornell.edu/iris/tutorial/dpm/index.html
Section 5 of this online tutorial provides a helpful introduction to appraisal and selection for data preservation