The UNESCO/PERSIST

Guidelines for the Selection of Digital Heritage for Long-Term Preservation

Edition II

By the UNESCO/PERSIST Content Task Force May 2021

Endorsed by:

ICA
International Council on Archives
Conseil International des Archives

IFLA
International Federation of Library Associations and Institutions

Contents

Foreword

The digital world is moving at lightning speed. Both digital content and digital technologies make significant advances every day, causing a serious challenge to heritage institutions and other information organisations to continue to select, preserve and access the documentary heritage output of the world. To ensure the long-term accessibility of significant digital heritage, identification and early preservation interventions are essential.

It has been five years since the first edition of the “UNESCO/PERSIST Guidelines for the selection of digital heritage for long-term preservation” was produced to support heritage and research institutions in selecting digital heritage for long-term sustainable preservation. In a digital world, five years is an eternity, and therefore a second edition of the Guidelines was called for.

The first edition of the Guidelines evolved from the work of the UNESCO/PERSIST project, which arose out of the Memory of the World Conference held in Vancouver, Canada, in 2012. The PERSIST project is a collaborative venture of UNESCO, the International Federation of Library Associations and Institutions (IFLA), the International Council on Archives (ICA), and other partners. In 2020, the PERSIST project was integrated into the work of the Preservation Subcommittee of the UNESCO Memory of the World Programme.

Impact

The first edition of the Guidelines was available in nine languages in print and electronic formats and was downloaded hundreds of times from the UNESCO and the IFLA websites. The text was mainly used by information practitioners in making selection decisions for preservation purposes. In addition, there were examples of the Guidelines being used in training workshops for digital information management. The Guidelines were intended to also help raise awareness among governments and civil society about the fleeting nature of digital information. It is hoped that as advocacy efforts by documentary heritage organisations take root, this second edition of the Guidelines will have wider distribution.

Relationship with UNESCO

UNESCO has been a staunch supporter of the past and current editions of the Guidelines, using them as a training tool as it raises awareness of digital information preservation and access in many parts of the developed and less developed world. In 2015, the UNESCO General Conference endorsed the “Recommendation concerning the preservation of, and access to, documentary heritage including in digital form.” Accordingly, every four years, UNESCO Member States are expected to report on their implementation of the 2015 Recommendation. These Guidelines are a useful resource that can assist UNESCO Member States in implementing the 2015 Recommendation.

The Writing Review Group

Many information professionals have been involved in creating this second edition of the Guidelines, and this during a very restricted time due to the COVID-19 pandemic. They have been very dedicated to accomplishing this task and are due much gratitude for sharing their expertise so enthusiastically.

Ingrid Parent, Chair (IFLA, University Librarian Emerita, UBC)
Claire McGuire, Secretary (Policy and Research Officer, IFLA)

Anthea Seles (Secretary General, International Council on Archives)
Davide Storti (Programme Specialist, Communication and Information Sector, UNESCO)
Fackson Banda (Programme Specialist, UNESCO Memory of the World Programme) (Ex officio)
Frédéric Blin (Head of Services and Collections, Bibliothèque nationale et universitaire, Strasbourg)
Gordon McKenna (Standards Manager, Collections Trust) (from 2021)
Ivy Lee (Head, Statutory Functions & Research, National Library Singapore) (from 2020)
Jenna Murdock Smith (Lead Archivist, Archives Branch, Library and Archives Canada)
Julia Chee (Deputy director of Collections, National Library Singapore) (to 2020)
Monika Hagedorn-Saupe (ICOM CIDOC Chair, Prussian Cultural Heritage Foundation) (to 2020)

Steve Knight (Programme Director, Preservation Research, National Library of New Zealand)
Winston Roberts (Senior Business Advisor, Office of the National Librarian, National Library of New Zealand)

We hope that you will find these Guidelines useful for your particular digital preservation needs.

Please send comments or questions to IFLA: ifla@ifla.org

Introduction

Heritage institutions — libraries, archives, and museums — traditionally bear the responsibility of preserving the intellectual and cultural resources produced by all of society. This important mission is now in jeopardy around the world due to the sheer volume, velocity and variety of information which is created and shared every day in digital form. In 2020, the International Data Corporation (IDC) Global DataSphere, which measures the amount of data created and consumed in the world each year, predicted that the amount of data created between 2020 and 2023 will be more than the data created over the previous 30 years (IDC, 2020). This exponential growth provides an increasing challenge to those working to preserve digitised and born-digital heritage for long-term accessibility.

Audience

These Guidelines are meant to be applied in all parts of the world and be used in all contexts. The Guidelines acknowledge that they cannot be too specific in their application because cultural heritage policies differ among countries, regions, and institutions. They provide a starting point for libraries, archives, museums, and other heritage institutions when developing their own selection policies for preservation.

These Guidelines are targeted primarily towards the practitioners in cultural institutions who make the day-to-day decisions about which digital materials are candidates for long-term accessibility. They support practitioners in reviewing existing policies, highlighting important issues to consider, and providing guidance in drafting institutional policies. Senior administrators in cultural institutions will also find these Guidelines useful in planning and prioritising strategies and budgets for managing their resources.

Key Challenges

Preserving the vast output of digital information is difficult not just because of its extent, but also due to its largely ephemeral nature. Digital media do not yet have the same longevity as physical objects, documents, and books, which often will survive for centuries. Digital file formats, storage media, and systems are ever evolving, jeopardising the future readability and integrity of digital heritage over a much shorter timeframe than paper and physical objects. Its availability for capture is fleeting. In addition, digital systems that are both tools for content creation and creations themselves are beginning to be recognised as digital content to be preserved.

The survival of digital heritage is much less assured than its traditional counterparts in our collections. Identification of significant digital information, items, or collections and early intervention are essential to ensuring long-term preservation. So too is addressing the economic burden of funding preservation — a challenge for national and institutional budgets, as well as a source of inequality between countries and communities.

It is clear that responsibility for the preservation of digital heritage goes beyond cultural institutions and will require the engagement and cooperation of both the public and private sectors, as well as content creators. The private sector will also face the challenge of preserving and ensuring access to its own digital information. This could be considered an opportunity for private-public sector cooperation and joint projects.

Diversity and Inclusion

The material for collection referred to by these Guidelines is intended to include digital content created by or about all ethnic, religious, gender, social, and political groups found in all regions of the world. The Content Task

Force recommends that archives, libraries, and museums consult and collaborate with underrepresented communities when making selection decisions to ensure that documentary heritage created by and about these communities is identified for long-term digital preservation.

The Review Group recognises that it is not best equipped to speak to the needs and perspectives of all marginalised groups in these Guidelines. Future initiatives that are led by those communities, such as Guidelines for the selection of Indigenous digital heritage for long-term preservation, are therefore welcomed and encouraged.[1]

Structure of the Guidelines

This second edition of the Guidelines is organised in two parts. *Part 1: The Impact of Digital Preservation on Selection Decisions* offers a concise look at the selection decisions which must be made by practitioners and others in order to preserve this new and significant part of the world's documentary heritage. Part 2 offers a deeper look at collecting in the digital environment.

These two parts are followed by several appendices, which provide more in-depth information related to several digital issues and their context.

Part 1: The Impact of Digital Preservation on Selection Decisions

Defining Digital Preservation

To define digital preservation, we refer to the definition provided by the Digital Preservation Coalition:

> *Digital Preservation refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary. Digital preservation… refers to all of the actions required to maintain access to digital materials beyond the limits of media failure or technological and organisational change. Those materials may be records created during the day-to-day business of an organisation; "born-digital" materials created for a specific purpose (e.g., teaching resources); or the products of digitisation projects* (Digital Preservation Handbook, 2015).

Digital preservation ensures long-term access to digital information across time, technologies, and semantic shift, and has long-term benefits, which may be social (evidence of government), cultural (national identity), and economic (use and re-use, innovation).

The digital shift has radically changed how cultural heritage is made, stored, preserved, disseminated, consumed, and monetised. Citizens' roles have also changed dramatically, shifting from passive observers to active participants and content producers with many new and exciting opportunities for engagement, creative use, and access (Tammaro, 2016).

The digital environment has created new forms of expression, ranging from web pages and interactive social media sites to private research databases, digital artworks, and online gaming environments. These materials blur boundaries and lines of responsibility, and they challenge past approaches to collecting.

[1] For introductory knowledge on Indigenous data governance, we recommend referencing the CARE Principles for Indigenous Data Governance: https://www.gida-global.org/care

Existing mandates and collecting policies often do not cover these new forms of digital heritage. Our collective neglect of these new forms raises the risk of creating large gaps in our cultural heritage for future generations.

As an example, although the value of individual posts on blogs or social media may be marginal, collectively they constitute a unique record of contemporary society — the discussions, thoughts, and achievements of billions of individuals. If preserved, this will represent an incomparable source of knowledge for future generations. Depending on the mandate of the institutions, focusing only on the "best" part of this output may introduce biases and prevent any analysis of contemporary digital creation as part of a whole.

Few, if any, heritage institutions have the resources, and in some cases the rights, to harvest and preserve *en bloc* this social output in digital form. *This is the paradox of selection in the digital age.* Selection is essential, as it is economically and technically impossible, and often legally prohibited, to collect all current digital heritage.

Selecting for long-term preservation will thus be a critical function of heritage institutions in the digital age and will need to be punctuated by practical, pragmatic, and collaborative approaches to capture a representative historical record.

How Does Collecting Digital Material Differ from Traditional Material?

Digital materials provide unique challenges, as they can often only be archived adequately in digital form. There is no non-digital equivalent that retains all the essential information and functionality it provides, and digital preservation is only useful if the material can be accessed at a later date (Digital Preservation Coalition, 2015).

Key differences between the collection of digital material versus traditional collecting include:

- The speed and complexity of change in the digital arena
- The need for intermediary tools and services to understand and manipulate the digital object
- The volume of information that must be reviewed for selection
- Continuing hardware and software obsolescence over time
- The speed and near invisibility of decay in digital content
- The scale at which digital collecting, storage, and preservation will increasingly need to be undertaken with the move from petabytes to exabytes over time.

See *Appendix 1: Digital vs Traditional Collecting* for a more detailed look at key similarities and differences between traditional and digital collecting.

Practitioners should also be aware that they must select and appraise materials across formats, as paper and digital often coexist in what can be referred to as a long *hybrid tail.* Selection practices therefore must be approached holistically, rather than as purely digital or purely paper/analogue. Other challenges arise relating to interlinking paper and digital formats to enable researchers a view in the round, but that is out of scope for these Guidelines.

Digital Material Types

Many traditional forms of cultural heritage now have digital equivalents, which often fit well within our existing practices and mandates. These include books, periodicals, government records, private correspondence, personal diaries, maps, photographs, film and sound recordings, artefacts, and works of art.

In addition to digital equivalents of analogue materials, digital heritage also takes the form of social media, Virtual Reality (VR) and Augmented reality (AR) material, digital art, web archiving, and many others.

While it is not possible to include an exhaustive list of all types of digital materials that one should consider, the Review Group felt it was important to examine several types of materials that may not be included in traditional content selection strategies.

Refer to the Appendices for a deeper look into selection of software source code, research data, social media, and artificial intelligence for long-term preservation.

Developing Selection Criteria

Existing institutional missions, mandates, and collection development policies for physical collections, in most cases, will provide the starting point and essential guidance for assessing and selecting digital heritage. These should now be adapted to embrace new forms of digital expression.

Evaluating and assessing digital heritage should be based on many of the same principles that underlie traditional selection, such as context and provenance, while acknowledging that some aspects of traditional collection may not transfer into the digital environment. Digital heritage also requires consideration for new issues of long-term accessibility, use, and preservation in making selection decisions.

Criteria for Selection

An institution should answer these questions by evaluating the relative **significance** of the digital heritage to its mandate and public; by assessing its **sustainability**, that is, the institution's capacity to preserve it for long-term access and use;[2] and by considering its **availability** in other heritage institutions, that is, its prospects for preservation elsewhere and the most appropriate institution or community group to preserve it.

The concepts of significance and sustainability in this environment must be evaluated in light of institutional mandates and resources. Availability looks outward to other institutions in the heritage community to assess the level of risk to the digital heritage's long-term survival. Particular attention must be paid to heritage that is at risk of being lost over the short to medium term, as well as heritage deemed to be of broad human significance, such as collections included on UNESCO's Memory of the World International Register.

Bearing in mind that every memory institution is unique in its mandate, collections policy, and resources, we offer below a series of steps and questions that can frame digital selection decisions. This approach can be scaled to the individual needs of institutions which are diverse in purpose and size. Even if not adopted, these steps can form the point of departure for an institutional discussion about the selection of digital heritage for long-term preservation.

Strategies for Collecting Digital Heritage

Heritage institutions must evolve their existing approaches to address the digital environment. It is likely that this will include adapting one or more of the following strategies or approaches.

Selection

Selection is used when heritage professionals — archivists, librarians, and curators — identify material for addition to their collections based on specific criteria. These criteria can vary widely depending on the type of institution, its collecting mandate, its resources, and the type and extent of material available for acquisition. The selection criteria are generally expressed or defined in a collecting or acquisition policy, and may be based on the following criteria (which may also be combined):

- *Function:* Documentary heritage institutions such as archives will select records for long-term value based on the work carried out by an institution or government. Example: Diplomatic wires are important reporting functions in Ministries of Foreign Affairs.

[2] It should be noted that institutions may choose to select significant records that may not be sustainable, with a view that infrastructures or technologies may be developed to allow their preservation or readability at a later point.

- *Subject/Topic*. An institution will focus on and attempt to document one or more subject areas. For example, all websites dedicated to a specific painter or locality, or a web crawl to document a specific event such as a political election or arts festival.

- *Creator/Provenance*. An institution will focus on particular creators of heritage or provenance. For example, an archive might acquire the digital records of authors of a particular region; or a museum might collect works of the artists of a particular movement.

- *Type/Format*. An institution might collect by the type or format of content (e.g., digital photography, music recordings, film, video games).

In some cases, institutions may choose to capture all the digital heritage material now and apply selection criteria at a later date, in a form of deferred selection. Below are two specific collection practices: comprehensive collecting and representative sampling.

Comprehensive Collection

Comprehensive collecting is used to acquire all of the material produced on a given subject area, time period, or geographic region. This approach requires significant institutional resources, or a narrow focus. Legal deposit of publications is perhaps the most familiar comprehensive approach, in which a national library attempts to collect the entire publishing output of the nation through a legal requirement for publishers to deposit copies of each publication they produce. Museums too might seek to gather all works created in a particular period.

Archives do employ comprehensive collecting practices in certain cases. For example, they might attempt to collect everything related to an influential public figure or particular nationally significant event.

Representative Sampling

Sampling is another approach used to identify material for long-term preservation. It is often used when an institution does not have the resources or capacity to collect comprehensively and when differentiating the material by specific selection criteria is problematic. In these cases, sampling offers a means of capturing a representative picture, making selection and preservation more manageable and less resource-intensive.

For example, a national library might perform regular crawls of a whole national web domain to preserve at different times a representative portrait of its national presence online. An archive might select government case files by using a sampling methodology, such as keeping only the most-documented cases, or those from a given year, or that begin with a given letter of the alphabet, or those determined by statistical analysis.[3]

Consider the Context

Before embarking on a project to select digital heritage, it can be useful to consider the overall contextual milieu at work. With increasingly vast amounts of information facing heritage professionals in a digital world, keeping attuned to context when researching and making selection decisions can save valuable time and resources and can lend a clarity to decision-making, which is at risk of becoming muddied in complex digital environments.

Examining the context of creation of the digital heritage, and the interrelationships between creators, activities, and systems can facilitate decision-making at a more strategic level. Considering context while navigating the decision tree below can be a useful guiding principle that can diminish tangents and will ensure that institutions acquire digital heritage with the metadata/provenance essential to preserving it and making it accessible.

As such, contextual considerations (“context checks”) have been included throughout the decision tree to assist with this analysis.

[3] Limitations with this approach have been identified by some government archives. See Cook, 1991.

Decision Tree for Selection in an Individual Institution

This approach consists of four steps, posed in a series of questions, to support consistent, evidence-based evaluation:

Step 1: Identification

Define the parameters of the project, if appropriate. Determine whether a *yes/no* decision is sufficient, or if a relative evaluation (high, medium, low) is required to compare materials.

Identify the material to be acquired or evaluated.

- What is its title, creator, provenance, extent, condition?
- Does the institution have the capacity and the rights to capture, preserve, and make available all the data/records/information?
- What kind and what amount of metadata is available? If the metadata is corrupted, how will this be addressed?[4]

Document your decision-making process in this step and keep the record. Ensure the record is kept up-to-date and accessible.

***Context check**:* Can identification be done at a higher level than the digital object itself? Are there functions or activities that can guide selection decisions?

For example, government archives may decide not to acquire any records created by administrative activities because they are sufficiently documented through high-level policy instruments and records supporting overall decision-making.

Step 2: Legal Framework

Start by asking the following questions to identify legal issues relating to the long-term preservation of your material:

- Does the institution have a legal obligation to preserve the material? Does the institution's mandate or policies on digital preservation and collections development require the preservation of this material?
- Have you considered potential obstacles to preservation, such as intellectual property rights or privacy restrictions?

If yes, preserve. If a positive decision is confirmed and no further selection is necessary, no further steps are required.

Document your decision-making process in this step and keep the record. Ensure the record is kept up-to-date and accessible.

Step 3: Application of Selection Criteria

If further selection is required, the institution can assess the material using three selection criteria: **significance**, **sustainability**, and **availability**. These criteria should be assessed in whichever order is most efficient or

[4] Metadata may be incomplete or corrupted because of migration, but this should not preclude a record from selection. It may be the institution retains the record, but supplements the metadata or finds an alternative mechanism to make it available while providing an explanation of the incompleteness of the metadata.

effective for the institution, generally beginning with the easiest criteria to evaluate and proceeding until a final decision is reached.

3(a): Significance

To determine the significance of an object, you will need to ask a series of questions regarding its authenticity, origin, reason for creation, and value to those who have created it. Start with the following:

- How closely do these values support and align with the institutional mission and mandate?
- Is the digital object born-digital or is it a surrogate of a physical item? Does the physical item still exist?
- Does it have significant social, cultural, historical, or artistic value for the community served by the institution, for the community that created the object, or for humanity as a whole?
- Does it have significant information, content, use, or research value?
- How will the institution's stakeholders (clients, sponsors, society) be affected if this digital item is not preserved?
- Is the documentary heritage material created by or about Indigenous people or an underrepresented community? If so, consultations with such communities should be considered as part of the selection process.

If the digital item is significant in relation to these criteria, consider preserving. Document your decision-making process in this step and keep the record. Ensure the record is kept up-to-date and accessible.

Context check: What information exists to identify activities/functions/creators that can aid with selection?

For example, are there any business analyses or administrative histories that can provide insight into why and how the digital heritage material was created? Can this help identify places/points of creation that may be crucial to document? What are the relationships among various activities/functions/creators that may reveal significant areas of value?

3(b): Sustainability

Before selecting digital heritage for long-term preservation, you must examine your capacity to provide sustainable access to the material.

The questions provided here are points of reflection. If an institution does not have the resources, capabilities, technical capacity, and/or budget to preserve and make material accessible, this does not necessarily mean that it should not acquire that material. There may need to be short-term, interim solutions developed to enable preservation and/or access. These decisions will need to be made on a case-by-case basis at the discretion of the institution and stakeholders involved.

It is important to remember that digital preservation is not simply a technical problem requiring only technical solutions. Curators, collection managers, information managers, conservators, and others responsible for the preservation of physical collections must understand the preservation needs of digital material. Ensuring that all jurisdictions are aligned in understanding the collecting aims of cultural organisations and the ethics that underpin the care of their digital collections will help meet the increasing challenge of building and sustaining these collections over time.

In every case, assessment of "sustainability" should be closely informed by "significance" and "context." Start by asking yourself the following questions:

- Does the institution have sufficient infrastructure to preserve digital heritage materials? If not, is there a contingency plan to ensure the sustainability of the material?
- Does the institution have technical capacity to comprehend, migrate and preserve the digital heritage?

- Are specific rights required to transfer or migrate the material to different file formats and/or physical carriers?
- Is sufficient metadata available to access and preserve the digital heritage?
- Is the format of the digital item suitable for preservation, or should it be migrated?[5]
- Can the institution make the material accessible for research, exhibition, or other use to meet the public's expectations?

Document your decision-making process in this step and keep the record. Ensure the record is kept up-to-date and accessible.

Context check: How was the digital heritage created and used over time? Are there links between systems or datasets that may alter the valuation of the digital heritage or create preservation concerns?

Understanding the intellectual control of the material over the course of its life can provide valuable clues for determining whether it is suitable for acquisition. For example, understanding the systems that an organisation used and the classification structures that were in place can be essential in making sense of why the digital heritage material looks the way it does.

A system may have been adapted to be used differently than the way for which it was originally designed. Thinking about the context of creation can allow an institution to select digital heritage material as early as possible in its life cycle — ideally prior to creation.

This can create opportunities for memory institutions to work with creators, advising on how their infrastructure and formats can be established from the outset to support long-term preservation.

Redundancy

Important digital heritage, including master files with associated metadata, should exist in multiple copies that are stored in at least two different physical locations. Heritage institutions can use a mix of on-site, off-site, and distributed cloud-based storage, but digital originals should be backed up in at least one other location. Storage sites should be chosen to diminish the risk of loss due to natural or human-caused disasters and economic or political crises.

Active management

Heritage institutions should actively manage their digital heritage assets to ensure their accessibility and integrity over the long term. Digital heritage should be preserved in open and well-documented file formats, without encryption, and at least lossless compression. This method is strongly recommended for heritage institutions in the active management of digital objects. Storage should use two or more different types of storage media, ranging from institutional servers to portable media (e.g., magnetic disk, optical media, or magnetic tape).

Systems failure over the long term can cause vital information loss to stored digital heritage. Many institutions guard against this by using a periodic media refresh, consisting of reading in the digital data, checking for errors using error correction techniques, and rewriting on new media. To avoid software failure, digital data owners often use standards-based protocols for access to data storage, where different storage sites are running different implementations of the storage software. Therefore, the integrity and reliability of data does not depend on the integrity and reliability of any single implementation.

3(c): Availability

[5] While there is an argument to be made for all formats to be preserved, this decision will depend on the practitioner's resources and infrastructure.

Consider the general availability of the digital heritage in other institutions in the heritage community or network. Start by asking the following:

- Is this institution the only one preserving this material, or are exact duplicate copies held by other institutions? Does it inform the fonds?[6] Or the overall institutional collection? Is it rare or unique, or is it widely duplicated?
- Where will it receive the most use or be of the most benefit to the public?
- Is it at risk at other institutions?
- Is this institution the most appropriate or best-placed to preserve and make accessible this digital heritage?
- Keep in mind that a certain amount of redundancy is necessary to secure proper preservation of digital heritage. See more on Redundancy in Appendix 6.

Availability, as with sustainability, should be informed by significance and context, especially if the records under consideration concern or were created by an underrepresented community.

Document your decision-making process in this step and keep the record. Ensure the record is kept up-to-date and accessible.

Context check: Are there important relationships among activities/functions/creators that can aid with selection?

Step 4: Decision

Compile and review all the records made during the process and make a decision based on the results from steps 1 to 3.

This approach is flexible; not every question will be applicable to each institution. Nor is the order of the criteria set in stone; for example, in some cases step 3(c) might be better assessed before steps 3(a) and 3(b), particularly if another institution is clearly more appropriate.

Document and record the rationale and justification for the evaluation or decision. This is vital, both for governance and to capture important information for potential reappraisal in the future. Prepare a written statement of the digital heritage's significance, context of creation, and its technical preservation issues, incorporating the answers to the questions in steps 1 to 3. The arguments behind the decision are often more important than the evaluation itself. A standard institutional evaluation form or appraisal document should be created to capture these arguments and be a record of the decision.

Despite these potential variances in application, following this approach should support heritage institutions in making better decisions in selecting digital material for long-term preservation.

Part 2: A Deeper Look at Collecting in the Digital Environment

When setting up a digital selection and preservation programme, it can be helpful to consider how collecting digital content may affect internal processes within one's institution. It is also beneficial to examine external issues, such as legal challenges and the role of partnerships, which both impact long-term digital preservation.

[6] Archival term, referring to the records aggregated from a given creator.

Digital Selection in Libraries, Museums, and Archives

In the digital world, the operating models and modes of serving the public being used by libraries, museums, and archives are changing rapidly. In this digital context, it is important that information professionals do not leave their traditional collection management activities to the technologists.

Digital is a paradigm shift and a clear challenge to the collection/preservation mandate, which requires a rethinking of how heritage institutions identify significance and assess value. This also emphasises the need for practitioners in libraries, archives, and museums to consider digital collecting together with traditional collection management processes, while acknowledging that each has unique considerations and approaches.

While some of the boundaries between libraries, archives, and museums are blurring in the digital age, there are still particular issues to keep in mind that are of specific relevance to each.

Libraries

Libraries will face the challenge of digital selection with respect to e-publications, harvesting of websites, and proprietary content in social media sites such as Facebook and YouTube. National libraries striving to build a comprehensive collection, often with a strong tradition of legal deposit, will have to adopt selection for more ephemeral publications in digital form. In the past, selection was done, in effect, by publishers who “curated” creative output through editorial choices that determined what would be published. In the democratised world of self-publishing and e-books, national libraries will have to modify past comprehensive approaches and adopt criteria to select for long-term preservation.

Although not all libraries hold cultural heritage collections that are recognised as significant in national or international frameworks, many are collectors of local or regional memory. The record of this memory, such as newspapers and audio-visual recordings, is increasingly likely to be born-digital material.

As libraries are mandated to serve their community’s needs, long-term accessibility and use are essential to preservation. Digital collections must be established with usability and means of discovery as central tenets. Selecting digital items for long-term preservation for some libraries may focus primarily on evaluating publications already in their collection, originally acquired for short-term use, rather than assessing new publications for acquisition.

At any size, most libraries will benefit from codifying their digital selection process and structuring their selection criteria on similar values shared among memory institutions. This will help libraries identify the types of digital material that may fall within their mandate to select for long-term preservation, avoiding both duplication and potential loss of material.

Museums

Museums with strong and well-developed collections of physical material culture generally acquire for permanent preservation and make collections development decisions in this context. This material culture is now increasingly digital — for example, machines which are driven by computer software, born-digital works of art, digital installation artworks, and digital documentation of archaeological sites. Research information related to the physical holdings of museums is also increasingly digital.

Further, museums are collecting digital heritage with a substantial physical component — for example, mobile phones, tablets, computers, and game consoles. These collection decisions in museums may be based on their mission statement. For example, a design museum may collect a “style icon” which is partly physical and partly digital, or a science museum will collect important hardware (and related software).

Digital heritage in museums thus can be divided into the following categories: born-digital items in the collection, digital or digitised information about the collection, and digital representations of physical artefacts in the collection (digital images or 3D scans for example). Due to this categorisation, museums normally should prioritise the first and second categories for long-term preservation. The second and third categories also include institution-generated administrative records.

The importance of metadata (information about physical and digital heritage) to museums cannot be overstated. This metadata includes contextual information created about the physical and digital heritage before it enters the museum and contextual information created during its life in the museum. The principle of provenance is also important to museums.

Archives

Archives focus on the importance of authenticity, provenance, and context in the appraisal of archival records for acquisition. The legal environment often dictates what digital information must be acquired by an archive and how, or if, it can be made accessible for public access and research. In some cases, openly licensed born-digital material can be archived massively, which configures new or evolved roles for the archives themselves.

Archives acquire original or unique records for permanent preservation and have traditionally relied on the passage of time between their creation (e.g., 20 years) and their acquisition to lend historical perspective in making selection decisions. However, the rapid obsolescence in digital formats, and the inability of record and data creators to maintain information, storage media, and system hardware and software systems, is collapsing the window of opportunity of selection. More archives are beginning to grapple with the need to collect born-digital records and data, social media, web software, artificial intelligence, and many other forms of significant data much earlier, which requires earlier engagement with records creators to influence the design and production of records and data of long-term significance. This has also required building closer working relationships with information technologists, data scientists and many others to ensure the integrity of the records and data creation process and the maintenance of information until it can be transferred. Early intervention by the community is needed because how records and data are created will influence their acquisition, preservation and access.

The capacity of archives varies significantly, but this does not mean that they should refrain from preserving digital materials (records, data, etc.) should they lack budget, capacity, and infrastructure. There are many maturity models that exist that help archives not only baseline their capabilities, but help them identify a practical roadmap forward considering the resources at their disposal.[7]

These factors which influence the selection environment are not necessarily exclusive to each of the library, museum, and archives communities. Indeed, there will almost certainly be some overlap. But reviewing the diversity of our communities helps illuminate the range of issues to be faced by institutions in identifying and selecting heritage for long-term preservation.

Partnerships for Digital Selection and Preservation

Memory and heritage institutions can benefit from sharing expertise and coordinate efforts on digital collection and preservation through participation in both national and international networks.

[7] Examples of digital preservation maturity models include NDSA *Levels of Digital Preservation* (https://ndsa.org/publications/levels-of-digital-preservation/) and DPC's *Rapid Assessment Model (*https://www.dpconline.org/digipres/dpc-ram*).*

International Cooperation

Taking part in joint international exchanges and experience-sharing is valuable for coordinating efforts on digital collection and preservation. Some countries will have already established national strategies and plans, which could serve as models elsewhere. Institutions with experience in managing digital collections can share their knowledge on issues such as legislation, standards and practices, digital selection and access systems, digital art conservation, preservation infrastructure, and security.

Archives, museums, and libraries work closely with their own international councils and federations for technical cooperation and exchanges. International civil society organisations such as ICA, IFLA, and ICOM, and intergovernmental organisations such as UNESCO should be consulted, as preservation of digital heritage is a global challenge, and these organisations can lead conversations in the international arena. These are also strong platforms on which to develop joint strategies.

The Role of National Institutions

National institutions of the library, archive, museum, and heritage sectors must play a vital role in providing leadership to government agencies and heritage communities on issues of digital selection and preservation.

In many countries, designated national institutions have legislation and policies related to collection development and management of cultural heritage. National libraries' legal deposit function enables collection of published works, and national archives are mandated to acquire official records of their governments. Museums have also embarked on collecting digital objects under their policies for contemporary collecting.

The scale and unique challenges of digital content make it natural for national institutions to take a leading role in digital content management. These institutions have existing standards, infrastructure, and systems that can be expanded further to support the rest of the heritage community, especially smaller archives, libraries, and galleries.

National institutions would also need to expand their network to garner expertise for a coordinated and collaborative plan to develop national selection, collection, and preservation strategies and standards. The network can include content creators, industry experts, heritage collectors, and key government agencies that deal with technology innovations, data protection, copyright, and digital media.

At the national institution level, it is important for libraries, archives, museums, and other heritage institutions to agree and define the roles and scope of national collections, including a cooperative collecting and preservation plan. It is recommended that selection of digital heritage content and its long-term management should be a national collaborative effort due to the challenges and scale of the digital content created within a country, and it may no longer be driven solely by a single institution. For example, if an institution does not have the requisite resources, capabilities, technical capacity and/or budget to preserve some material, it might work in partnership with other institutions with more capacity at a national or regional level that can help facilitate recommendations.

The Importance of Advocacy

At the time of writing, ongoing restraints due to COVID-19 are a defining challenge, but general budget cuts and other difficulties will surely continue to be concerns. Acknowledging the many challenges facing the public sector, it is more important than ever to raise awareness of the cultural value of collecting digital heritage, both among the public and decision-makers.

The aim of this advocacy should be to emphasise the benefits derived from long-term access to digital material. For example, this can highlight the legal mandate of making cultural heritage accessible to the public, or it can speak to the economic, intellectual, and scientific benefits brought about through access to this digital content.

Institution-level Advocacy

At the local or national level, general advocacy may be relatively neutral and take the form of public relations, advertising campaigns, awareness-raising, and professional seminars.

Perhaps the best advocacy that a memory institution can employ is running informative public events, featuring well-informed and engaging staff. This can have the long-term benefit of gaining public support, demystifying the institution, dispelling misunderstanding, and communicating how this work benefits society at large.

Partnerships with Civil Society
Cooperation between practitioners and civil society organisations may be effective in this regard, as it is not always appropriate for national institutions to take a leading role in advocacy and lobbying efforts. International organisations such as IFLA, ICA, ICOM, and intergovernmental organisations such as UNESCO, are actively engaging in advocacy campaigns relating to the preservation of and access to heritage, including in digital form.

Successful advocacy can influence the direction of intergovernmental debates, can result in effective technical standards, or can gain support and funding for professional activities.

The Impact of the Legal Environment

The legal environment has important implications for the selection, preservation, and availability/accessibility of digital heritage. International and national laws vary widely, but they normally regulate the capture, dissemination, duplication, access, and use of digital heritage. However, the internet transcends territorial boundaries, and therefore often makes it difficult to determine which laws apply and to identify rights-holders.

Legal deposit in libraries enables the capture, preservation and accessibility of digital published material. (See IFLA's checklist on legal deposit[8] for more issues to consider around legal deposit.) Meanwhile, archives are governed either by **archival legislation** or by **acquisition policies.** Archival legislation provides archives with the mandate to collect and manage digital public records emanating from ministries or government departments and units. An acquisition policy more often governs the collecting, preservation and accessibility of digital materials in private sector organisations. By contrast, the collections of a museum may not be governed by law. National museums tend to work more often within a more defined legal framework. However, the situation varies from country to country or can vary by the governance of the museum.

A strong, enabling legislative framework is an essential condition for a successful digital selection and preservation programme. In its 2003 Guidelines for the Preservation of Digital Heritage, UNESCO noted that:

"As a key element of national preservation policy, archive legislation and legal or voluntary deposit in libraries, archives, museums and other public repositories should embrace the digital heritage. Copyright and related rights legislation should allow preservation processes to be undertaken legally by such institutions."

[8] https://www.ifla.org/publications/node/93470

It is therefore important that organisations undertaking a digital preservation programme ensure that the legislative context in which they operate enables them to capture digital content for safekeeping.

To consider:

Do legal deposit laws and archival legislation cover digital content, and if so, of what type? Do the current laws impede the acquisition of nationally significant documentary heritage? If so, how can institutions ensure the selection and preservation?

Do legal deposit laws restrict collecting to publicly accessible works, or are there possibilities to collect or request works that are paywalled or otherwise restricted?

What rules are there around the treatment of works which subsequently become unavailable or restricted online?

Other Legal Considerations: Copyright and Privacy

Copyright legislation, save for specific exceptions and limitations, may prohibit the making of copies. This raises new issues in a digital environment in which duplication may be necessary for long-term preservation. In addition, digital materials are often software-dependent for search and retrieval, and this software may also be protected by copyright.

Some countries have enacted laws to prohibit circumvention of technological protection measures used to prevent copying and redistribution, thereby hindering preservation and impeding future legitimate access to digital heritage. Other countries have legal provisions which enable them to circumvent technological protection measures for the purposes of preservation.

There are jurisdictions that have **“fair use” provisions,** which some organisations use for collecting material deemed to be in the public domain (e.g., publicly available websites). This means that this material can be made available to users without seeking prior approval.

Other jurisdictions use an **“opt out”** in the same way, requiring an objection voiced by the copyright owner before the material is made available.

By understanding the relevant legal implications of digital preservation in their local jurisdiction, practitioners can be better equipped to select digital heritage for long-term preservation and develop a legal liability framework accordingly.

Privacy laws may also impact the availability and accessibility of digital information that contains personal data. This will be dependent on the jurisdiction, but “Right to be Forgotten” laws, for example, may impact what cultural heritage organisations can acquire, preserve and make available.

The lack of international consistency in copyright laws, and lack of clarity on how to cooperate across borders, can have an unwanted impact on efforts to preserve works (for example, through digital preservation networks), store them (for example, in the cloud) and give access, all affecting content selection decisions.

In order to support the preservation of digital content, a key long-term goal is international action to enable safeguarding of materials and giving access to heritage practitioners, alongside progress at the national level to ensure that the choices of digital collections managers are guided, as far as possible, by professional concerns alone.

To consider: Does legal deposit legislation give institutions clear permission to carry out preservation by copying, in any necessary form?

Can institutions ignore contract terms and circumvent technological protection measures?

What rules are there on giving access to collected digital works?

What are the provisions for preservation of personal data? Do privacy or data protection laws exist? How do they impact on what can be acquired, preserved and made available?

Conclusion

As digital technology opens a world of possibility for human expression and ingenuity, it also establishes the unique challenge of selecting what remains accessible. Likewise, as the technology used to create these expressions evolves, so must the selection approaches, along with preservation and access methodologies.

The interventions and decisions of practitioners of today will influence the memory for tomorrow. This is an enormous responsibility that memory and heritage institutions face in partnership with other national-level institutions and lawmakers, as well as with international networks and civil society organisations.

There must be strategies put into place to select this digital material for long-term preservation, to codify the selection and collecting process, and to document how decisions were made. Practitioners are invited to take the selection criteria and background provided here as a guide and adapt it to fit their context. The appendices that follow will give a deeper dive into some key concepts that could provide even more insight when creating a selection strategy.

The Guidelines have been written in acknowledgement of the ever-changing landscape of information creation that defines our time. Take these Guidelines as a piece of a larger, international effort towards preserving and providing access to the world's digital heritage — an effort which surely will continue to evolve long into the future.

Appendix 1: Digital vs. Traditional Collecting

This chapter takes a deeper look into the similarities and differences between traditional and digital collecting. This includes a risk management example to demonstrate how these differences affect preservation in practical terms.

Similarities

Digital collecting is not "a technical problem requiring only technical solutions. Rather, it is a social, cultural, and organisational problem, just as traditional conceptual and ethical approaches to physical [collections] were derived from social and cultural concerns" (Slade, S., Pearson, D., & Knight, S., 2019).

As with physical collecting, digital collecting is "characterised by provenance, context, chain of custody, and thorough documentation over time. The digital preservation environment can be thought of in the same way as the storage environment for physical collections. Preservation storage for digital collections consists not only of the room and storage equipment but also the organisation of the collection objects, methods for accurately locating them, security provided to them, policies about their preservation and care, the environment within which

they are placed, and methods for monitoring the environment to ensure that any risks to their preservation are controlled" (Slade, S., Pearson, D., & Knight, S., 2019).

Risks to both physical and digital collections are similar, but differ in how they are manifested. (See Table 1.)

Differences

The effort and investment required to safeguard digital material should not be underestimated. The endeavour will be ongoing and intensive, and will continue to extend as scale, format, and accessibility issues become more challenging.

Issues specific to digital collecting include:

- Obsolete formats (e.g., WordStar)
- Digital material on non-supported physical carriers (e.g., floppy disks, Betacam)
- Cloud-only publishing models
- Modern platform jurisdictions — born-digital material on overseas platforms not subject to local collecting legislation (e.g., Facebook, YouTube, Instagram, Twitter)
- Lack of capability/capacity across the knowledge system
- Outdated models of intellectual property management (e.g., copyright) requiring new access, licensing and use/re-use models
- Lack of a coherent response to long-term issues of protection of digital assets for innovation, use and re-use, and reproducibility (i.e., being able to validate the integrity of research findings)
- Undermanaged digital information in the public sector limiting opportunities to maximise the value of government information.
- Digital material that is composed of or necessitates separate items (files) that are not necessarily available (legally) or archived

Table 1 — A risk management example of the similarities between digital collecting and traditional collecting (Pearson, 2012).

Risk	Definition	Digital example	Physical example
Parameter-based risks	A criterion identified by staff to indicate a preservation risk	Video encoded within a problematic codec	The identification of film as cellulose nitrate
Exceptions risk	The value of a monitored parameter is outside a set of acceptable values	A file in a particular format fails validation	The relative humidity levels within a store are outside the agreed levels
Change risks	A change in the status of a monitored parameter for content	The confidence in format identification for a particular file has changed, or a checksum has failed	The fading of colours in a colour photograph or watercolour due to prolonged exposure to visible light
Conflict risks	Conflicting values for the parameter are reported by one or more tools	A file format identification returns conflicting values	Different temperature and relative humidity readings for the same collection store are obtained from the building management system and independent environmental monitoring

Unknown value risks	Undetermined values for defined parameters	There are unknown values for file format and version	An upholstered chair is showing signs of deterioration; the cause cannot be identified and may be from material used during the construction of the chair and no longer accessible
Access support risks	Changes in the level of support that affect the organisation's ability to access content in accordance with its preservation plan	A significant reduction in the availability of supporting software for a particular file format	A surfactant used for washing textiles is no longer available due to changes in environmental guidelines legislation
Content based risks	Characteristics of content that may not be identifiable from metadata	The presence of deprecated HTML tags	Earlier painting beneath the surface of an oil painting; no information about this associated in the provenance files for the painting in this collection and no knowledge that it exists

Appendix 2: A Closer Look at Software Source Code

Software permeates our personal and social lives. It embodies a vast part of the technological knowledge that powers our industry, supports modern research, mediates access to digital content, and fuels innovation (UNESCO et al, 2019).

Software is written by humans in the form of *software source code*. It is a valuable and unique form of knowledge that, besides being readily translated into machine-executable form, is also "written for humans to read" (H. Abelson et al, 1985) and "provides a view into the mind of the designer" (Shustek, 2006). Despite being ubiquitous, software source code is often disregarded as an accessory to "executable" software programmes, which run in personal computers or through internet-based cloud services. While the rise of free and open-source software paved the way for the development of publicly accessible code hosting platforms, a significant part of software source code remains inaccessible, in the hands of private companies, or simply unshared for various reasons.

The Paris Call on Software Source Code as Heritage for sustainable development states that software source code is essential to preserving this precious technical, scientific, and cultural heritage over the long term (UNESCO & INRIA, 2009). In other words, software source code is today a peculiar born-digital artefact that embodies the knowledge and the effort that goes into the making of digital tools and creations shaping today's digital world, including research and archiving. In a digital driven world, software source code is an essential output of research and should be considered, along with research publications and research data, a pillar of open science.

Until very recently with the opening of the Software Heritage archive (softwareheritage.org), software source code lacked both the place to permanently store source code as a digital object and the mechanisms to efficiently

identify, store, and reference source code digital objects — e.g., through intrinsic persistent identifiers (Software Heritage, 2020).

Selected readings

Bussi, L., Di Cosmo, R., Montangero, C., Scatena G. (2019). *The Software Heritage Acquisition Process*. UNESCO. Accessed: https://unesdoc.unesco.org/ark:/48223/pf0000371017/PDF/371017eng.pdf.multi.

Di Cosmo, R. (2020). *How to use Software Heritage for archiving and referencing your source code: guidelines and walkthrough.* Accessed: https://annex.softwareheritage.org/public/guidelines/archive-research-software.pdf.

Appendix 3: A Closer Look at Research Data

Scientific research is the first element contributing to the progress of knowledge and therefore to the progress of humanity. Countries and international organisations throughout the world spend an important part of their annual budget to sustain, foster, and improve scientific research activities. International collaborations between scientists are common, as are international large-scale initiatives in fields such as space research, nuclear technology, and medicine. All these research activities create huge amounts of raw data that may need several years, if not decades, to decipher, compile, compare, and analyse before allowing any scientific conclusion — thus creating new knowledge.

Open research data allows other researchers, besides the creator of the data, to reuse it. Uses for open research data include validating or invalidating first results and introducing alternative tools or methods to draw new information and knowledge. For research to be validated, scientific publications must be able to refer to data that is available and confrontable through a new lens.

The concept of FAIR data (*Findable, Accessible, Interoperable, Reusable*) is now becoming the common ground on which research is and will be made. So, what if research data — that cost millions, if not billions, to produce — is no longer technically available because of a lack of preservation?

The preservation of digital research data is essential to science and knowledge. The definition of *Data Management Plans* has become the norm for every research project that seeks funding from its institution or external agencies. However, there is no definition of what needs to be preserved, how it needs to be preserved, or for how long.

From an archival perspective, there is an increasing demand for litigation-related research and requests under access to information and privacy research that may come with an expectation that information be released digitally in a timely fashion. With the plethora of versions and copies created in a digital environment, this can create a significant administrative burden on memory institutions.

Critical Questions for Selection

The parts of research data that are worthy of long-term preservation as heritage material must be defined through a thorough methodological and professional critical analysis. This is similar to the methodology used in archives. Some questions may include:

- What is the scientific quality of the data? How was it created? By whom? How open is it?
- Are there privacy issues linked to the data, and in that case, how long are these privacy issues valid?
- Can the experience that created the data be easily reproduced, for a minimal cost?
- Was the data created through a non-duplicable experience or at an expensive cost?

States, funding agencies, and research organisations must invest in the preservation of research data, under the FAIR principles.

Collaboration among institutions, on a national or an international level, allows the construction of sustainable infrastructures to manage the storage and preservation of ever-growing amounts of research data sets. Post-custodial models where data remains with a creator institution could also be considered. In some cases, the creator institution may be able to preserve this data more effectively themselves, rather than sending it to a heritage institution.

Librarians, archivists, information professionals, data technicians, engineers, computer scientists, and researchers all have roles to play, thus enlarging the scope of professions responsible for the preservation of our scientific heritage.

Appendix 4: A Closer Look at Social Media Appraisal and Selection

What is social media?

Social media content is generated by users on an interactive web service. The content can be images, texts, audio, or whatever is accepted by the platform.

How do you identify social media records?

Every social media record has an identification number or reference to identify a new item on the social media website. The elements from the structure and design that remain the same for all the users of the social website are not social media records but need to be preserved with a web archival system like other websites.

There are many important questions that need to be considered when looking to acquire social media records. Namely, is your institution best placed to scrape and preserve this material?

How do you work with service providers? How do you capture social media records? What are the things you need to think about from a technical point of view?

We work with Application Programming Interfaces (APIs) and downloading tools to obtain the files with all the records. There are many definitions of what constitutes an API; the most helpful are those put forward by Wikipedia and the Society of American Archivists.

Social media records can be captured with APIs in most cases, but there are some instances where more advanced techniques such as scraping are needed.

Strategies and tool types:

- **Web scraping/crawling**: A web scraper or crawler is browsing software that downloads data, like a web browser obtaining raw data but not formatted by default.
- **API orders**: Obtaining data with terminal console commands that send queries to the application programming interface (API) and receive formatted data with specific parameters.
- **Exporting server/profile data**: Making formal petitions to the social website services and receiving all the data of our profiles and those of the users we manage.
-

The resources below will provide more information on these types of strategies and tools that can be employed to capture social media.

Things to think about before capturing social media records:

- Are you the best organization to capture this information? If so, are there partners you need to involve?

- How much do you want to capture? Or how much can you capture? Do you have enough free space in your servers? Or external hard drives?
- Do you have a preservation strategy for all these social media records?

Social challenges of collecting social media

While some social media platforms are open by default, others are used more privately, restricted to a certain audience. It is worth examining how collecting this content may impact individuals, families, or communities, and how the intended audience informs collecting decisions. Some questions to consider:

- Do users see social media as public or personal information?
- Do the users of social media platforms see themselves as "publishers" and expect their content to be in scope for collecting by a state agency?
- Can there be a multi-layered framework regarding the ethics of social media collecting? Or is there a need for different approaches to different platforms?
- How much is understood regarding the different ways people use social media?
- When is it necessary to ask for permission to collect and when is it not? Is it viable to seek retroactive permission? What kind of access to these collections can be provided in each case?

Legal challenges of collecting social media

- Should all social media be treated the same or are there different approaches depending on how a social media platform is being used within society?
- How is it possible to balance the mandate to collect with the need to protect privacy? How is this done while ensuring the institution is a trusted repository for personal data?
- What is the local jurisdiction's position regarding the "right to be forgotten"'?
- How can "mixed-jurisdiction" digital content (such as material from Twitter) be managed?

Selected readings

- Corrado, E. M., Moulaison, H. L. (2017). *Digital Preservation for Libraries, Archives and Museums*. (2nd Ed.). Rowman & Littlefield.
- Russell, M. A., Klassen, M. (2018). *Mining the Social Web*. O'Reilly Media. (3rd Ed.)

Appendix 5: A Closer Look at Artificial Intelligence

Introduction

The use of artificial intelligence (AI) is becoming a mainstay in public and private sector organisations. It is seen as an effective and efficient means to analyse large amounts of structured data (i.e., data sets) and unstructured data (e.g., word processing documents, presentations, audio-visual content). AI allows organisations to gain insights from their data which would be difficult to obtain from a human-only process. It enables the analysis of large amounts of data to make policy decisions, develop medical treatments, produce advertising campaigns, and much more. The outputs from these tools are facilitating the decision-making process in various sectors, and that results in records that need to be documented and captured.

This section will start off with a definition of artificial intelligence, followed by an examination of its components that need to be evaluated and considered during the appraisal and selection process. That is followed by a discussion of several factors that affect the acquisition of AI records, such as the impact of public-private partnership and the capacity and resources of documentary heritage institutions to acquire these records. The section will conclude with a discussion of AI in appraisal and selection processes.

This section is a first attempt to delineate a process around the appraisal and selection of AI records. It will evolve as documentary heritage practitioners gain more insights and experiences.

Definition

Artificial intelligence can be defined in many ways and there is no single agreed definition, but for the purposes of this chapter we will define it as: “a branch of computer science dealing with the simulation of intelligent behaviour in computers; the capability of a machine to imitate intelligent human behaviour” (Marr, 2018). The concept of AI can then be broken down into two broad groupings: Supervised and Unsupervised.

Supervised algorithms or supervised machine-learning is where an algorithm is trained to identify patterns using labelled training data; for example, people’s names or places. It is then provided with raw untagged data to assess its levels of accuracy (“Precision and Recall,” 2021).

Unsupervised algorithms or unsupervised machine-learning is where an algorithm is given untagged data and it then identifies patterns, using built-in statistical probabilities (i.e., Bayesian inference). This process, unlike that of supervised machine-learning, is done with minimal human intervention (“Unsupervised learning,” 2021).

There are many different subcategories and types of supervised and unsupervised machine learning, such as deep machine learning, latent semantic indexing, and natural language processing. The purpose of this chapter is not to define and describe all of them, but rather to give the reader an overview.

How Do You Appraise and Select AI?

There are a number of considerations when appraising and selecting artificial intelligence. AI, whether supervised or unsupervised, is more than just the outputs of the algorithms. There are many different components to this type of record that may need to be assessed and captured, such as logs, data (structured and unstructured), iteration of code, and final code along with outputs (i.e., visualisations). Practitioners will also need to assess the context of creation, such as the creator’s reason for creating and using an algorithm for decision-making, as this will affect what may fall in or out of scope for acquisition. Finally, there are also the legal considerations that may impact what may be acquired. This is addressed in detail in an earlier chapter, but for this section we will consider the consequences of public-private partnerships, especially with regard to government usage of AI in policy decisions.

i) When is AI a record?

As mentioned, there may be legal considerations that need to be taken into account when trying to determine whether AI constitutes a record to be acquired, but other questions practitioners should ask to assess its long-term value would be:

- Did the outputs from artificial intelligence impact policy development and the application of government policy? Does this affect citizens? Does it affect their ability to assert and defend their rights?
- Did the outputs from artificial intelligence change an organisation’s mission?
- Did the use of artificial intelligence influence a key organisation project? Or did it affect how a decision was made?
- Did the use of artificial intelligence mark a turn in how an organisation or government made its decisions?
- Were there multiple entities or organisations involved in the development and training of AI? If so, who owns the AI code? Who owns the training data? Who owns the outputs? Are they the same entity?

Essentially, if the response to the first four bullet points is “yes” then AI constitutes a record of enduring historical value that should be acquired by a documentary heritage organisation. The last question attempts to surface the

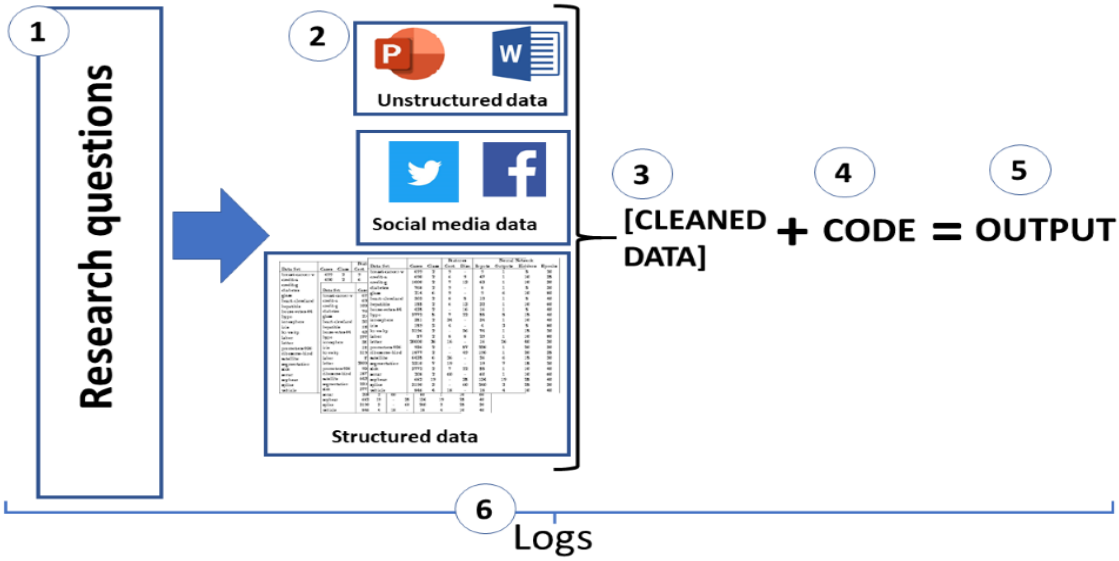
challenges when there is more than one entity involved in the creation, training, and usage of AI. The last bullet will be discussed in greater detail in section *iii) Public-private partnerships,* but it might have a significant impact on the selection process.

The scope of what may be acquired can only be assessed by understanding the different components of AI. Nevertheless, the breadth and volume of the components identified for transfer may exceed the capacity and capability of documentary heritage institutions. The constraints faced by these institutions may need innovative and out-of-the box solutions to ensure the capture and preservation of these records. We will explore some ideas in the section *iv) Capacity of heritage institutions.*

ii) Components of artificial intelligence

An AI record is not simply the output of an algorithm, because to understand how it arrived at the output, you need the data, logbooks and code. Figure 1.1 outlines the different components of a simple AI record, meaning that there is only one piece of code along with its supporting documentation (logs and data) that need to be captured. This figure presumes that the organisation using the algorithm has control over the entire process from the development of research questions, to collation and cleaning of the data, to development and iteration of the code to arrive at the output.

Figure 1.1 Components of a simple artificial intelligence record



1) Research questions
Any initiative using artificial intelligence has research questions that it needs to answer. These questions will be the basis for identifying relevant data, and will inform decisions related to data cleaning and iterations in the development of the code to arrive at the final result. The research questions will need to be understood by a practitioner at the beginning of the appraisal and selection process, as they may inform what will fall in and out of scope for transfer.

2) Data
Different types of data (structured, unstructured, and social media) will be identified and brought together in order to train the algorithm. Structured data consists of data in tabular form (e.g., Excel spreadsheet) whereas unstructured data is data that does not have a predefined model (e.g., word processing document or presentation). Social media qualifies as both structured and unstructured and as such stands as a separate data category (Price, 2013). The selection of data should be documented in a log, and it acts as an audit trail for the AI development and training process. The log should contain information about the data such as its source, content, date range, and sample set.

3) Cleaned Data
All the data coming from different sources will need to be aggregated and cleaned. The term *cleaning* entails that data that is not relevant to research questions or that may need to be aligned with other data (e.g., changing

centimetres to metres) that needs to be fed into the algorithm. The data cleaning phase is needed to facilitate comparison and analysis of information by the algorithm. All the decisions made during the data cleaning phase should be documented in logbooks and will enable the reproducibility of results by researchers and computer scientists in order to confirm the findings and decisions that may result from the AI's analysis (Mackenzie, 2019).

4) Code
Code is a computational and mathematical representation of the response to the research questions. It enables the algorithm to interpret and analyse data to arrive at an "answer" or response. The code is developed iteratively by developers or researchers to arrive at the best possible response to the research questions. As part of this process, the accuracy (precision and recall) should be assessed to ensure the integrity of the algorithm's output.

The code development process can be documented in the logbooks and/or in a piece of software designed to audit the code development process. Again, it is important that this process be documented to support the reproducibility of the algorithm's results, verifying the integrity of the findings and, by extension, the decisions that were made using the output.

5) Output
The output is the result of the computational processes applied by the algorithm and can be rendered in different ways, such as visualisations or statistical probabilities.

iii) Impact of public-private partnerships

In many instances, AI initiatives, especially in the public sector, consist of partnerships between a public body and a private company specialising in AI. This has many implications for the selection process and is related to the discussion around intellectual property and copyright earlier in this document. Essentially, practitioners need to determine:

- Who owns the code or algorithm? If it is owned by the company, then how do you document the AI output?
- Who owns the training data set? And the subsequent data fed into the AI?
- Who owns the output of the AI (e.g., data set or visualisation)?

Ostensibly, public bodies need to be accountable for the decisions they make using AI, and whilst intellectual property considerations should be taken into account prior to implementation, that may not be the case for policy decisions. This does pose ethical issues for accountability, transparency and the future public record, and workarounds may be needed to document the use of AI in decision-making. It may be worthwhile for practitioners to examine "algorithmic accountability" manifestos such as the Montreal Declaration and the ACM Declaration.

iv) Capacity of heritage institutions

The issue of capacity (infrastructure, resources, and personnel) is partly addressed in the appraisal and selection decision tree earlier in this chapter, in particular under "sustainability," but practitioners should be more swayed by the significance of AI. Sustainability will be an ongoing issue, and the question of how to ingest, preserve and make these materials needs to be explored in greater depth by the documentary heritage community. As such, whilst capacity will need to be assessed, the real issue will be what interim solutions need to be put in place for the short term while the community explores more sustainable long-term approaches.

Other considerations

Use of AI in appraisal and selection of born-digital records

AI is a record deserving of selection and preservation, but given the volume and complexity of digital materials that need to be assessed for selection in the future, it is not possible for documentary heritage practitioners to unilaterally apply traditional appraisal and selection techniques. There will need to be some form of automation moving forward, and the profession needs to assess the strengths and weaknesses of these technologies to understand where best to apply them and when. There are documentary heritage institutions that are experimenting with these technologies for appraisal and selection.

Further Reading: AI in Appraisal and Selection

- Caplan, R., Donovan, J., Hanson, L., and Matthews, J. (2018). Algorithmic Accountability: A Primer. *Data and Society*. https://datasociety.net/wp-content/uploads/2018/04/Data_Society_Algorithmic_Accountability_Primer_FINAL-4.pdf

- Chumtong, J., Kaldewey, D. (2017). Beyond the Google NGRAM Viewer: Bibliographic Databases and Journal Archives As Tools for Quantitative Analysis of Scientific and Meta-Scientific Concepts. *FIW Working Paper No 8.* https://www.fiw.uni-bonn.de/publikationen/FIWWorkingPaper/fiw-working-paper-no.-8

- Engin, Z., Treleaven, P. (2018, August). Algorithmic Government: Automating Public Services and Supporting Civil Servants in using Data Science Technologies. *The British Computer Society*. https://academic.oup.com/comjnl/advance-article/doi/10.1093/comjnl/bxy082/5070384

- Ertzscheid, O. (2017). *L'appétit des géants: pouvoir des algorithmes, ambitions des plateformes*. Paris: C&F.

- Information Privacy Commissioner. (2017). *Big Data, Artificial Intelligence, Machine Learning and Data Protection*. London: ICO. https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf

- LeSueur, A. (2016). Robot Government: Automated Decision-Making and its Implications for Parliament [Draft chapter for publication in *Parliament: Legislation and Accountability*. Oxford, UK: Hart Publishing. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2668201

- National Archives and Records Administration. (2020, October). *Cognitive Technologies White Paper: Records Management Implications for the Internet of Things, Robotic Process Automation, Machine Learning and Artificial Intelligence.* Washington, D.C. https://www.archives.gov/files/records-mgmt/policy/nara-cognitive-technologies-whitepaper.pdf

- The National Archives UK. (2016). *The Application of Technology Assisted Review to Born-Digital Records Transfers, Inquiries and Beyond*. (2016). London. http://www.nationalarchives.gov.uk/documents/technology-assisted-review-to-born-digital-records-transfer.pdf

- O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishing.

- Padilla, T., Allen, L., Potvin, S., Roke Russey, E., Varner, S. (2017, March). Collections as Data. https://doi.org/10.17605/OSF.IO/MX6UK.

- Rolan, G., Humphries, G., Jeffrey, L., Samaras, E., Antsoupova, T., Stuart K. (2018, November). More Human than Human? Artificial intelligence in the archive. *Archives and Manuscripts, 47*(2), 179-203.

- World Wide Web Foundation. (2017). *Algorithmic Accountability: Applying the Concept to Different Country Contexts*. https://webfoundation.org/docs/2017/07/Algorithms_Report_WF.pdf

Appendix 6: Management of Metadata

Metadata

Metadata is usually defined as “data about data,” which, though accurate, is not very precise. In heritage institutions, the required metadata should be considered as any information (in digital or physical form) that is essential to ensuring that the digital material being preserved is, and remains, accessible, intelligible, and usable over time. Metadata provides the institution with the information required to access and preserve digital heritage into the future.

Heritage institutions generally preserve three key types of metadata associated with digital heritage that are crucial to long-term preservation:

- Structural (required for the technical capacity to read digital content)
- Descriptive (containing bibliographic, archival, or museum contextual information, which can be system-generated or created by heritage professionals, content creators, and/or users)
- Administrative (documenting the management of a digital object while in its collection)

If the digital heritage is the “content,” then the metadata provides the “context.”

There are five basic functional requirements for digital metadata:

Identification: The metadata must allow each digital object to be identified uniquely and unambiguously. This will usually require a globally unique identifier to be assigned to each item.

Location: The metadata must allow each digital object to be located and retrieved. The long-term validity of this location data needs to be ensured so that items are not lost when systems are migrated or updated.

Description: A description of each digital object is needed to facilitate recall and interpretation. Descriptive metadata falls into two categories: data about content and data about context. Data about the content of an item can often be re-created by examination and consultation. It is nonetheless useful as a finding aid for resource discovery. Data about context — where, when and by whom an item was created, what it was used for, its place in relation to a general corpus of material — is much more difficult to re-create once lost.

Readability: Metadata about the structure, format, and encoding of digital objects is needed to ensure that they remain legible over time. This functional requirement is particularly important for digital objects, as they cannot be read without mediating technology. This metadata should identify the relevant standards and provide references to the technical documentation, authority files, and other related material needed for a complete rendering of the digital resource. Care needs to be taken to ensure that all the multiple layers of a digital object can be interpreted — from the encapsulating file format to the representation and codification of the data itself.

Rights management: Rights, conditions of use, and restrictions applicable to each digital item need to be recorded in the metadata. This metadata should identify the applicable laws and conventions and provide references to relevant legal documentation, contracts, etc., as well as the rights holders.

Storage of Metadata

Many digital file formats allow metadata to be embedded within the file itself. This has the advantage of ensuring that the data and metadata remain linked. However, metadata also needs to be stored independently from the digital resource that it describes; this is essential to meeting the functional requirements set out above. An encoded digital item, for example, cannot be read if the code is only to be found embedded in the item itself.

Metametadata

Some data about the source of the metadata and how it was compiled is needed to establish its reliability and authenticity. For future retrieval and understanding of the digital information, contextualisation is essential.

This data can include:

- When was the metadata compiled and by whom?
- Was the metadata harvested automatically or manually?
- What tools and techniques were used?

Appendix 7: References

Abelson, H., G. J. S., Sussman, J. (1985). *Structure and Interpretation of Computer Programs*. The MIT Press.

Cook, T. (1991). Many are Called but Few are Chosen. *Archivaria, 32.* *https://archivaria.ca/index.php/archivaria/article/view/11759/12709*

Digital Preservation Coalition. (2015). *Digital Preservation Handbook* (2nd Ed.). https://www.dpconline.org/handbook

Digital Preservation Coalition. (n.d.). *What is Digital Preservation?* https://www.dpconline.org/digipres/what-is-digipres

Digital Strategy for the Library of Congress. (n.d.) The Library of Congress. Retrieved December 3, 2020, from https://www.loc.gov/digital-strategy

Harvey, R. (2006). Instalment on Appraisal and Selection. *DCC Digital Curation Manual* (version 1). *https://www.era.lib.ed.ac.uk/bitstream/handle/1842/3331/Harvey%20appraisal-and-selection.pdf?sequence=1*

International Data Cooperation. (2020). *IDC's Global DataSphere Forecast Shows Continued Steady Growth in the Creation and Consumption of Data*. https://www.idc.com/getdoc.jsp?containerId=prUS46286020

Koerbin, P., Webb, C., & Pearson, D. (2013). 'Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections. *Magazine of Digital Library Research*, 19(1/2). http://www.dlib.org/dlib/january13/webb/01webb.html

MacKenzie, R. J. (2019). Repeatability vs. Reproducibility. *Technology Networks.* https://www.technologynetworks.com/informatics/articles/repeatability-vs-reproducibility-317157

Marr, B. (2018). The Key Definitions of Artificial Intelligence (AI) That Explain its Importance. *Forbes.* Retrieved January 28, 2021, from https://www.forbes.com/sites/bernardmarr/2018/02/14/the-key-definitions-of-artificial-intelligence-ai-that-explain-its-importance/?sh=dd7c4a64f5d8

National Library of Australia. (2003). *Guidelines for the Preservation of Digital Heritage*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000130071.

Niu, J. (2012). An Overview of Web Archiving. *D-Lib Magazine*, 18(3/4). http://www.dlib.org/dlib/march12/niu/03niu1.print.html

Niu, J. (2014). Appraisal and Selection for Digital Curation. *International Journal of Digital Curation*, 9(2), 65-82. http://www.ijdc.net/index.php/ijdc/article/view/9.2.65

Pearson, D. (2012, March 26-27). *The adventures of Digi: Ideas, requirements and reality* [Conference Presentation]. Future Perfect 2012, Wellington, New Zealand. https://www.nla.gov.au/content/the-adventures-of-digi-ideas-requirements-and-reality

Precision and Recall. (2021). In *Wikipedia*. https://en.wikipedia.org/wiki/Precision_and_recall

Price, G. (2013). The Difference Between Structured and Unstructured Data in Social Media. *Social Media Today*. https://www.socialmediatoday.com/content/difference-between-structured-and-unstructured-data-social-media

Shustek, L. J. (2006). What Should We Collect to Preserve the History of Software? *IEEE Annals of the History of Computing*, 28(4), 110–112.

Slade, S., Pearson, D., & Knight, S. (2019). An introduction to digital preservation. In L. Elkin, & C.A. Norris (Eds.), *Preventive Conservation: Collection Storage* (pp. 809-829). New York: Society for the Preservation of Natural History; American Institute for Conservation of Historic and Artistic Works; Smithsonian Institution; The George Washington University Museum Studies Program.

Software Heritage. (2020). *SoftWare Heritage persistent IDentifiers (SWHIDs)* (version 1.5). https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html

Tammaro, A. (2016). Heritage curation in the digital age: Professional challenges and opportunities. *International Information & Literary Review*, 48, 122-128.

UNESCO. (2003). *Charter on the Preservation of Digital Heritage.* http://portal.unesco.org/en/ev.php-URL_ID=17721&URL_DO=DO_TOPIC&URL_SECTION=201.html

UNESCO & INRIA. (2019). *Paris Call — Software Source Code as Heritage.* https://unesdoc.unesco.org/ark:/48223/pf0000366715

UNESCO & University of British Columbia. (2012, September 26-28). *UNESCO/UBC Vancouver Declaration*. The Memory of the World in the Digital Age: Digitization and Preservation, Vancouver, British Columbia, Canada. http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/pdf/mow/unesco_ubc_vancouver_declaration_en.pdf

UNESCO, University of Pisa, & INRIA. (2019). *The Software Heritage Acquisition Process*. https://unesdoc.unesco.org/ark:/48223/pf0000371017/PDF/371017eng.pdf.multi

Unsupervised learning. (2021), In *Wikipedia*. https://en.wikipedia.org/wiki/Unsupervised_learning

Werf, T. van der, & Werf, B. van der. (2014). *The paradox of Selection in the Digital Age* [Paper presentation]. UNESCO open session, IFLA WLIC 2014, Lyon, France. http://library.ifla.org/id/eprint/1042