The UNESCO/PERSIST DRAFT Guidelines for the selection of digital content for long-term digital preservation

A draft by the UNESCO/PERSIST Content Task Force

August 2015

Table of Contents

Introduction

At the Memory of the World Conference in Vancouver, Canada (September 2012) international experts agreed that there is a pressing need to establish a roadmap for solutions, agreements and policies, that ensure long term access and trustworthy preservation of digital information. The roadmap should address issues like open government, open data, open access and electronic government. In order to establish an effective roadmap, governments, social organizations and the IT-industry will have to work together. As a result of this meeting the *UBC/UNESCO Vancouver Declaration The Memory of the World in the Digital Age: Digitization and Preservation* was launched.

The agreement in Vancouver was the starting point of the PERSIST Project, which stands for Platform to Enhance the Sustainability of the Information Society Transglobally and was launched at an international conference in The Hague on 5 and 6 December 2013. The PERSIST Project has three task forces (content selection, technology, and policy) addressing different areas and challenges in long term digital preservation.

These draft Guidelines are presented in August 2015 by the Selection Guidelines Committee of the Content Task Force. The aim of these Guidelines is to provide an overarching starting point for libraries, archives, museums and other heritage institutions when drafting their policies on the selection of digital content for long-term sustainable digital preservation. The Guidelines are aimed at institutions and professionals on every level and in every region of the world in order to give an overview of existing materials for selection, as well as provide some guidance when drafting their own selection policy.

1. Sampling and collection strategies

Heritage institutions are traditionally entitled to preserve the intellectual and cultural production of their own nation, whatever their form or their medium may be. This long-standing mission is now challenged by the sheer amount of content which is produced and published every day in digital form, especially on the internet. Digital technologies have indeed dramatically eased the creation and production of original content and reduced their costs. National web domains range from thousands to millions of websites; on which millions to billions of files are posted, updated or deleted every day. Government administrations and private companies produce an unprecedented mass of digital records. What is not saved today by heritage or other institutions is at risk of being lost in the near future.

This content easily generated by users poses specific challenges for heritage institutions. Indeed, collecting and preserving this production in its entirety is not an achievable goal anymore. Therefore, institutions are sometimes tempted to focus on the cultural artifacts that would be the digital equivalents of the resources they previously gathered in physical form. This restriction poses a problem: even though the individual value of each blog, webpage or public expression on a social media is often disputable, this set of resources constitutes an incomparable testimony of contemporary societies, the opinions, discussions and achievements of billions of citizens. They will represent an un-rivalled source for future humanists and social scientists. Focusing only on the "best" part of the digital creation would

create strong bias and prevent any analysis of the contemporary digital production as a whole.

Heritage institutions are thus facing a conundrum: selection is necessary, as it is legally, economically and technically impossible to gather all current digital production. They need to continue collecting, in digital form, what they were gathering on a physical media. However they also need to preserve tracks of citizen expressions and activities in digital form. To tackle this challenge, these institutions generally rely on mixed approaches:

- **Comprehensive collection of digital equivalents of physical artifacts**. Institutions often try to ensure the collection continuity for types of content that were already gathered on physical forms: libraries tend to maintain a comprehensive legal deposit for e-books, e-periodicals, or audiovisual documents published by major producers; museums seek to gather all works produced in a particular time period, etc. However, even for this kind of content, comprehensiveness may be challenging: the rise of self-publishing leads to a dramatic increase in the number of available e-books; most people produce more emails than paper documents, for example.

- **Selection of online content according to specific criteria**. Digital content (websites, blogs, some parts of social network, or offline content such as some emails, office document etc.) is individually selected by heritage professionals – archivists, librarians and museum curators. This content may be selected according to different criteria, notably:
  o The topic of the content (e.g. all websites dedicated to a specific painter). Event-based web crawls are a specific case of topic-oriented selection: institutions will collect all digital documents (notably websites, blogs and social network) related to a political election, a sport or cultural event or others.
  o The producer of the content (an official institution or a company for archives; a writer for libraries; an artist for museums etc.)
  o The type of content (text, images, sound, audio, interactive documents, video games and others).

- **Representative samples of whole web domains**. In order to keep the memory of the internet production of their own countries, some institutions – generally national libraries – perform regular crawls of whole national domains, such as .dk, .uk or .fr. These web crawls often span over millions of websites. However, they cannot be considered comprehensive: many websites are not identified by archiving robots, frequently changing content is lost, and the depth of harvesting may not be enough to cover all pages of a single website. However, this kind of approach helps to automatically preserve a representative sample of a national web domain.

2. The role of Institutions

In many countries, the National Library and National Archives have legal deposit legislation to enable them to collect valuable commercially or privately published materials, and public records. Most of the current legislation worldwide covers only physical materials. In today's

increasingly digital world, born digital materials are being created and published in such large quantities that national libraries and national archives find it necessary to include digital materials in their legal deposit legislation, to varying degrees of success.

As it requires immense effort and resources to acquire and harvest digitally produced materials through the various digital channels and platforms, it is essential that proper policies and systems are set up to manage the end-to-end process to ensure that both physical and digital heritage materials are systematically collected and preserved in perpetuity for access and use.

In reality, as most national libraries, national archives as well as museums do not have sufficient resources and expertise to collect especially born digital materials in time before they disappear, it is critically important that other partners, i.e. governments, academic institutions, researchers/research centres, non-profit and private sector organisations take on digital stewardship roles and responsibilities. The national institutions could jointly define the standards and processes of collecting, organising and preserving while various other institutions can complement these standards and processes. To ensure that this happens, a network of digital stewardship organisations could be set up by the government, library, archives, museums and the community of practice. This could then define the collection and preservation responsibilities, and share those responsibilities.

3. Shared nature of long term preservation

Long-time preservation of digital documents is a problem that libraries, museums and archives are facing every day. Whatever the origin, content or format of digital information, they face the same challenge: to ensure that information will survive long term changes in storage media, devices, and data formats.

Electronic documents do not have the same longevity as physical materials, stored in libraries, archives and museums, many of which have been preserved for centuries. Traditional text and images were transmitted on paper. Paper objects (books, newspapers, prints, maps, music scores, manuscripts, etc.) are housed in closed storage facilities to avoid damages through excessive handling. Nevertheless they can be regarded as relatively safe. The electronic items, both digitized and born digital, are more fragile and susceptible to different factors that may jeopardize their longevity and integrity. However digital objects can be copied without severe quality degradation and transmitted remotely. They do still face quality loss challenges, such as random bit errors or damages to entire disks. Therefore the lifetime of digital media is much less predictable than the lifetime of traditional media.

The main goal of memory institutions in the field of digital preservation is to reduce the risk of loss and to retain cultural heritage in digital form and ensure their legibility, interoperability, availability, and authenticity over a long period of time. Preservation of electronic data is done not by one single technique, but by regarding all possible scenarios which may cause loss of the material. An efficient system for long-term storage and archiving should solve the problems which might cause losses or damages.

The general approach in libraries, archives and museums is to realize that some losses are inevitable, but reliability can be improved by redundancy, active management and metadata preservation.

Redundancy

It is generally recommended that important digital data, including master files with all associated metadata should be stored on at least two kinds of media, in different physical locations. To avoid serious data losses most memory institutions decide to back-up their data off-site. This approach has limitations, because the data storages could be threatened by natural or man-made disasters. Therefore, one recommended solution would be using distributed multi-site redundant storage by spreading information among multiple physical locations and regularly checking the storage integrity. The remote sites should be selected carefully to diminish probability of loss due to economic, environmental or political failure.

Active management

Often the choice for secondary storage is removable media such as optical disks or magnetic tapes. It is common to place the most used files on magnetic disk, less frequently used files on less expensive and slower optical media and very infrequently accessed files on magnetic tape, as offline storage.

The material should be guaranteeing its availability and integrity over a long period of time. The general recommendation is using open and well documented file formats, without encryption or at least lossless compression using open formats. Data stored in multiple locations have to be available and consistent for the desired lifetime of the documents. This method is strongly recommended by memory institutions in the active management of electronic items.

In systems designed to provide long-term access to archived electronic documents a basic recommendation is avoiding failure which causes information loss. A common technique is periodic media refresh comprising reading in the digital data, checking for errors using error correction techniques and rewriting on new media. It is also important to ensure that the stored data can be read in the future. Many repositories store multiple representations of the same document or use file formats of archival quality having specifications precise enough to build a credible interpreter or make conversion. 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. Thanks to such solutions the integrity and reliability of data does not depend on the integrity and reliability of any single implementation.

Metadata preservation

In digital repositories there are three main types of metadata associated with digital objects: descriptive (containing bibliographic description of the digital object), structural (gathering components of the complex object and administrative), and supporting resource management within collection.

Metadata can identify the digital object, document the technical processes connected with digital preservation, specify information on rights management and establish the authenticity

of digital content. Preservation of metadata is an important component of the digital preservation strategy in the institution.

The general goal of metadata preservation is supporting and facilitating the long-term retention of digital information, especially providing staff with appropriate knowledge to take actions in order to maintain the digital resources for a long time and ensuring that the content of all archived objects can be accessible and interpreted in future.

Schemas of metadata preservation which contain detailed description of all types of metadata associated with the archived resources should be prepared for every digital repository of heritage institutions.

4. Legal implications

The use of information technology in capturing or converting information facilitates easy and convenient access to and exchange of digital content, whether they are text, databases, pictures, maps, music, films, games or others. The dissemination, copying, access and use of digital resources are regulated by law and vary among countries.

Legal issues involving the protection of intellectual property rights, privacy and confidentiality, etc. affect what and how digital heritage can be preserved, and if and when they can be made accessible to the public. In some jurisdictions, legal deposit laws applicable mostly to national libraries or similar institutions provide for ownership or custodianship of publications but do not address adequately the need for preservation and copying, particularly for digital content with multi-right holders. Copyright laws, save for a few exceptions for libraries and archives, prohibit the making of copies and use unless consent is obtained from the copyright holders. Moreover, digital materials are often software dependent for search, retrieval and other functionalities but software vendors with competing business considerations may not be relied on for long-term preservation. Some countries have enacted laws on technology protection measures to prevent circumvention of technological devices used to control unauthorized copying and redistribution, which could inhibit preservation measures and impede future access to digital heritage. Most personal archives contain personal data that are subject to data protection or privacy laws to ensure fair and secure processing of personal data for narrowly defined purposes unless consent is given by the data subjects. Depositors or donors may also wish to make specific agreements regarding the confidentiality of certain materials placed in a library, archives or museum.

To meet legal requirements, before digital materials are selected as heritage for preservation, access and use, one must determine if permission from copyright or other rights holders is required and to whom permission should be sought. This applies not only to the content of the material but also to the associated software. Dialogue with the rights holders needs to be established for negotiation and clearance of rights by means of contract, licence, and grant conditions or through statutory provisions, particularly when preservation, access and use of the material may require migration into new forms or emulation of the original computing environment.

As legislation in many countries are not designed for or have not been updated with the digital environment in mind, legal restrictions on the dissemination, copy and use of digital materials have considerably influenced their selection as heritage for preservation, access and use. Organizations may choose to digitize materials which they have rights or easy clearance of rights much sooner than those with restrictions attached. Preservation initiatives may deliberately include only those that are right-free to avoid possible legal complications. This could have a negative impact on the protection and availability of digital heritage for posterity. Thus cooperation from right holders must be actively pursued for the purpose of protecting, preserving and accessing digital heritage, especially when the rights holders do not want or do not have the capacity to take up long term preservation responsibilities. International and domestic legislation aiming at overcoming legal restrictions and providing for clear roles and responsibilities of parties in the selection and preservation of digital heritage for public access and use should be introduced as soon as possible as the internet has largely transcended the territorial and legislative boundaries of information exchange, making it difficult to identify the right holders and the law from which jurisdiction should apply.

5. Recommendations concerning metadata for digital preservation

Metadata is usually defined as data about data, which is accurate but unhelpful. In the current context, the metadata required for digital preservation can be considered as any data, in addition to the material to be preserved, that is needed to ensure that stored digital items remain accessible, intelligible and usable over time.

There are five basic functional requirements for digital metadata.
- **Identification**: the metadata must allow each digital item 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 item 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 digital items is needed to facilitate recall and interpretation. Descriptive metadata falls into two categories: data about content and 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 recreate once lost.
- **Readability**: metadata about the structure, format and encoding of digital resources is needed to ensure that they remain legible over time. This functional requirement is particularly important for digital resources 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 resource 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 meet 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; when was the metadata compiled and by whom? was the metadata harvested automatically or manually? What tools and techniques were used?

6. Assessing the long-term value of Digital Heritage

Selecting digital heritage for long-term preservation is an important function of libraries, archives, and museums in the digital age. Heritage institutions will need to assess the long-term value or significance of digital material available for inclusion in their collections. Digital heritage, which must be migrated over time to new physical carriers and ever-evolving file formats, is compelling a rethinking of how the heritage community assesses value.

What digital heritage should be preserved for the long-term? How should libraries, archives, and museums select, identify, and prioritize the digital material of most value to their stakeholders? How will they ensure that the authenticity of digital heritage is maintained over time through multiple migrations and format changes? An institution can answer these questions by evaluating the relative significance of the digital heritage being considered to its mandate and public, and by assessing its capacity to preserve it for long-term access and use. Only by careful consideration of the relative value of the material and the risks at the time of acquisition will institutions make informed decisions concerning its long-term preservation.

Although every heritage institution is unique in its mandate, collections policy, and resources, we have prepared a decision tree[1] (see the Appendix 1) that is flexible and scalable to institutions that are diverse in size and purpose. It can be adapted to the needs of individual institutions to support evidence-based evaluation decisions.

---

[1] This is based on http://cultureelerfgoed.nl/sites/default/files/publications/assessing-museum-collections.pdf

Appendix 1: Decision Tree for Selecting Content

Bearing in mind that every heritage institution is unique in its mandate, collections policy, and resources, the following approach is designed to be flexible and scalable to institutions that are diverse in size and purpose. This approach consists of six steps, framed in a series of questions, to support consistent, evidence-based evaluation and preservation decisions.

Step 1
Identify the material to be acquired or evaluated. What is its title, provenance, extent? Define the parameters of the project, if appropriate. Is a simple yes/no decision sufficient, or is relative evaluation (high, medium, low, or a numerical valuing?) required to compare it against other material?

Step 2
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?
If yes, preserve. A positive decision is confirmed, and no further assessment is required.

Step 3
Does the long-term value of the digital heritage justify its preservation? Does it have significant social, cultural, historical, or artistic value for the community served by the institution? Does it have significant information, content, use, exhibitions, or research value? How closely do these values support and align with the institutional mission and mandate? Does the material's provenance, rarity, or representativeness affect these values? How will the institution's stakeholders (clients, sponsors, society) be affected if this digital heritage is not preserved?
If the digital heritage is not significant in relation to the institution's mandate, do not preserve.

Step 4
Is this institution the only one preserving this material, or are exact duplicate copies held by other institutions? Is this institution the most appropriate or best-placed to preserve this digital heritage?
If the answer is no, perhaps the digital heritage should be preserved by another institution. But other factors in steps 3 and 5, should still be weighed against this answer.

Step 5
Does the institution have sufficient budget and resources to preserve the digital heritage material over the long-term? Does the institution have technical capacity to read, migrate and preserve the digital heritage? Can it ensure that its authenticity is maintained over time? Are specific rights required to transfer or migrate the material to different file formats and physical carriers? Is sufficient metadata available to access and preserve the digital heritage? Can the institution make it accessible for research, exhibition, or other use to meet its public's expectations?
If the answers are no, do not preserve.

Step 6

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 re-evaluation. Prepare a written statement of the digital heritage's significance and its technical preservation issues, incorporating the answers to the questions in steps 3 to 5. 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.

This approach is flexible, not every question will be applicable to each institution. Nor is the order of the steps set in stone; for example, in some cases step 4 might be better assessed after step 5. But following this approach will support heritage institutions in making good decisions in selecting digital material for long-term preservation.

Appendix 2: Definitions of terms

The definitions are taken from sources listen in Appendix 3 under *References relating to Definition of terms.*

Authenticity of digital heritage refers to the trustworthiness of a record or an item, i.e. the quality of being what it purports to be, either as an original object or as a reliable copy derived by fully documented processes from an original.

Content selection refers to the process of deciding which items or materials are worthy for preservation through evaluating their significance and lasting cultural, scientific, evidential or other value as well as the feasibility of preservation and access based on defined principles, policies, procedures and standards.

Digital heritage means heritage which is made up of computer-based materials, whether born digital or digitized from other format, which emanates from different communities, industries, sectors and regions and requires active preservation approaches to ensure its authenticity, accessibility and usability through time.

Heritage refers to legacy from the past, what we live with today, and what should be passed from generation to generation because of its significance and value.

Metadata means information that describes, explains, locates or otherwise makes it easier to understand, retrieve, use, manage, control or preserve an information resource or item through time.

   For generic selection criteria:
Value (relate to what to preserve including value to the preserving organization, value to present and future users, social significance, uniqueness and the risk of not preserving, etc. Authenticity, integrity and metadata come into play here for they affect value):
- Physical conditions, technical and format issues and legal rights (relate to the feasibility and costs of long-term preservation as well as how to preserve and future use); and
- Resources and policy of the preserving organization (unique to each organization).

For method and procedure:
- Who to select (professional and stakeholders) and the workflow/steps;
- Timing of selection (early selection), whether reappraisal is required; and
- Mode of selection (apply the selection criteria to measure value and other elements; use random sampling to ensure representativeness/diversity; and/or conduct risk analysis to ascertain uniqueness and other value against obsolescence, damage and costs, etc.).

Appendix 3: References

The list of references is a selection and in no way meant to be exhaustive.

References relating to the Introduction:

"UBC/UNESCO Vancouver Declaration The Memory of the World in the Digital Age: Digitization and Preservation", Canada, 2012

Van der Werf, T. and B., 'The paradox of selection in the digital age', 2014

Helmus, W., "Survey on selection and collecting strategies of born digital heritage - best practices and guidelines", 2015

References relating to Definitions of terms:

Charter on the Preservation of Digital Heritage, UNESCO, 2003

Digital Preservation Handbook Definitions and Concepts, Digital Preservation Coalition, 2008

InterPARES 2 Project Dictionary, 2015

ISO 15489-1 Information and documentation – Records management – Part 1: General, ISO 2001

ISO 16175-1:2010(E) Information and documentation – Principles and functional requirements for records in digital office environments – Part 1: Overview and statements of principles, ISO 2010

A Glossary of Archival and Records Terminology, Society of American Archivists, 2005

General reference:

Harvey, R. "Appraisal and selection" DCC Digital Curation Manual (2006) (which aims to offer a generic selection framework)

Niu Jinfang "Appraisal and Selection for Digital Curation", International Journal of Digital Curation, 2014 Vol. 9 Issue 2, 65-82 (which gives a good summary and includes many relevant references)

References relating to sampling and selection strategies:

Normative documents:

Charter on the Preservation of Digital Heritage, UNESCO, 2003

IFLA Statement on Legal Deposit, IFLA, 2011

CENL/FEP Statement on the Implementation of (Statutory and Voluntary) Deposit Schemes for Non-Print Publications, FEP and CENL, 2012

Information and documentation — Statistics and Quality Indicators for Web Archiving (draft available at: http://netpreserve.org/sites/default/files/resources/SO_TR_14873__E__2012-10-02_DRAFT.pdf), ISO, 2013 (This ISO technical report provides a terminology for web archiving techniques; it also presents the issues of web archiving for national libraries and proposes measures, assessment and characterization criteria for resources harvested online.)

Overviews of digital legal deposit and web archiving initiatives worldwide

Digital Legal Deposit, an IPA [International Publishers Association] Special Report, 2014

Jinfang Niu, "An Overview of Web Archiving", D-Lib Magazine, 2012, Volume 18, Number 3/4

Specific examples and use cases

National library, Australia

Colin Webb, David Pearson, Paul Koerbin, "Oh, you wanted us to preserve that?!' Statements of Preservation Intent for the National Library of Australia's Digital Collections", D-Lib Magazine, 2013, Volume 19, Number 1/2

National library, France

Sylvie Bonnel, Clément Oury, "Selecting websites in an encyclopaedic national library: a shared collection policy for internet legal deposit at the BnF", 2014

National library, United Kingdom

Caroline Brazier, "http://library.ifla.org/222/1/198-brazier-en.pdf.", 2013

National archive, United Kingdom

Amanda Spencer, John Sheridan, David Thomas, "UK Government Web Continuity: Persisting Access through Aligning Infrastructures", The International Journal of Digital Curation Issue 1, Volume 4, 2009

Museum (New York art resources consortium),

"NYARC Reframing Collections for a Digital Age Report from Consultant No. 2", 2012, USA

References relating to the role of institutions:

Noonan, D.W. "Digital preservation policy framework: A case study', Educause Review 2014,

2015 National Agenda for Digital Stewardship, 2014

British Library digital preservation strategy 2013 – 2016, British Library, 2013

References relating to the shared nature of long-term preservation:

Charter on the Preservation of Digital Heritage, UNESCO, 2003

Library of Congress Digital Preservation Webpage

British Library digital preservation strategy 2013 – 2016, British Library, 2013

Digital Preservation Handbook Definitions and Concepts, Digital Preservation Coalition, 2008

The State of Digital Preservation: An International Perspective, Conference proceedings, Washington D.C., April 24-25, 2002. Washington, DC, 2002

Verheul, I. "Networking for Digital Preservation: Current Practice in 15 National Libraries", München, 2006

References relating to legal implications:

Charter on the Preservation of Digital Heritage, UNESCO, 2003
Jules Lariviere, "Guideline for Legal Deposit Legislation", UNESCO, 2000
Kenneth Crews, "Study on Copyright Limitations and Exceptions for Libraries and Archives", Standing Committee on Copyright and Related Rights, World Intellectual Property Organization, Twenty-ninth Session, Geneva, December 8 to 12, 2014
Workbook on Digital Private Papers, Paradigm, 2005-7

References relating to metadata:

van Ballegooie, Marlene and Duff, Wendy "Archival Metadata", DCC Digital Curation Manual, 2006
Cunningham, Adrian, "Recent Developments in Standards for Archival Description and Metadata" IASA Journal no 16., 2000
ISO 15489-1 Information and documentation – Records management – Part 1: General, ISO 2001
ISAD(G), "General International Standard Archival Description 2nd Edition", Ottawa: ICA, 2000

References relating to the long-term value of Digital Heritage:

Netherlands, Cultural Heritage Agency, "Assessing Museum Collections: Collection Valuation in Six Steps", 2014
Canada, Canadian Heritage Information Network, Digital Preservation Toolkit, 2013
Catherine A. Bailey, "Past Imperfect? Reflections on the Evolution of Canadian Federal Government Records Appraisal", Archivaria 75, 2013,
InterPARES 2 Project, "Preserver Guidelines - Preserving Digital Records: Guidelines for Organizations", 2007