Community Digital Preservation Survey 2014

Assessing the current state-of-the-art in digital preservation practices

Guidance notes v1.0

Table of Contents

Introduction

Background and Aims

The community survey aims to build an evidence base about adoption of tools and approaches in the community. The survey will ask about your organisation, your core activities and collections, and about the technology you use.

Why do I need a guide?

In order to keep the survey form as uncluttered and readable as possible, detailed information about each question has been provided in this guide. This provides extra information about the intention and format of the question, so you can prepare your responses in advance, and it also links to reference material about the services and software we are asking about where this is available.

Who should fill out the survey?

It is unlikely that any one person working for medium or larger size organisations will have all information required to complete the survey alone. Please feel free to work as a group, or to gather information from the people you work with in your preferred manner. Our only requirement is to have one response per institution so that the data can be analysed easily.

What will be done with the information?

We will publish our findings, fully anonymised so that individual institutions cannot be identified, to demonstrate shared priorities across the community.

The data from the survey will be available for others to reuse, including the selected data from our member survey issued in October. Published findings will focus on sector-wide trends, to provide an analysis of the adoption of tools and approaches which can act as a reference for members and the broader community in developing their own practice.

Do I need to answer all the questions?

Certain questions are mandatory for all respondents, while some are relevant only to content-holding institutions, and some are entirely optional. We have not used complex logic in the online survey to determine which questions you should answer (other than the mandatory questions) so please do answer all questions which are relevant to you. This is particularly important for content holding institutions - even if the question appears to be optional the response may still be important. These are indicated in the guidance document.

The survey structure

About your organisation

This section captures background information about your institution, including your staff capacity, your digital activities, and the variety and scale of digital content that you hold. We have worked with many different people to define the vocabulary but we appreciate it may not always match local terminology. We have aimed to make the categories as broadly useful as possible across a range of different contexts, we hope you find them helpful.

Q1: Name of organisation? [*mandatory; free-text*]

Any published results will be fully anonymised so any information you supply will be in complete confidence.

Q2: Type of organisation? [*mandatory; drop-down select*]

The type of organisation, selected from a pre-populated list. This is to enable an analysis of trends across different kinds of organisation.

- *Academic/research library*
- *Art gallery*
- *Data centre*
- *Funding body (e.g. research council or foundation)*
- *Government department*
- *Institutional archive*
- *Institutional library*
- *Museum*
- *National archive*
- *National library*
- *Regional archive*
- *Regional library*
- *Representative body (e.g. community/membership organisation)*
- *Research unit (e.g. university department)*
- *Service or infrastructure provider*
- *Software product or solution provider*
- *Other [please specify]*

Q3: Where is your organisation based? [*mandatory; drop-down select*]

Please tell us in which country your organisation is based. This will help to identify trends geographically.

Q4: Number of staff in your organisation? [*mandatory; drop-down select]*

The overall number of staff in your organisation or unit (as chosen in Q2), to enable an understanding of staffing capacities and provide information about the ratio between digital preservation staffing within the institution as a whole (using the answer to Q5).

- *1-5*
- *6-20*
- *21 - 50*
- *51 - 200*
- *200 - 1000*
- *1000+*

Q5: Number of staff in your digital preservation team by role? [*mandatory; multiple free-text*]

The roles and number of staff who deliver your digital preservation programmes. A list of roles is provided, against which you can put figures for Full Time Equivalent (FTE). If staff only work part time on digital preservation please count only that proportion. For example: one full-time member of staff or two half-time members of staff would both be 1.0FTE; one full-time and one half-time would be 1.5FTE.

- *Cataloguer or metadata analyst*
- *Collection development or digital librarian*
- *Conservator or e-conservator*
- *Content or workflow analyst*
- *Data manager*
- *Digital archivist or curator*
- *Legal (IPR) specialist*
- *Manager or administrator*
- *Outreach or user-support specialist*
- *Records manager*
- *Repository or collections manager*
- *Researcher*
- *Software developer or programmer*
- *System administrator*
- *Usability or user experience specialist*

Q6: Do you have any other roles in your digital preservation team? [*optional; free-text*]

Please tell us if there are any roles missing in Q5, which form part of your digital preservation team. We have tried to cover a range of roles, but recognise that digital preservation departments are structured differently across organisations and sectors.

Q7: What are your core activities in relation to your digital preservation programmes, and how do you deliver them? [*for content-holding institutions; multi-select checkboxes*]

We would like to understand whether these are carried out **in-house**, **out-sourced** to an external supplier, or rely on **community solutions**. This will enable an analysis of how digital preservation programmes are delivered and highlight trends across the sector for use of external suppliers or requirements for community solutions.

- *Digital forensics*
- *Digitisation*
- *Emulation*
- *Format identification*
- *Format migration*
- *Format validation*
- *Metadata creation/extraction*
- *Policy development*
- *Preservation education and outreach*
- *Preservation planning*
- *Preservation research*
- *Software development or maintenance*
- *Storage or bit preservation (incl. backups, fixity, security etc.)*
- *Technology watch*
- *Web harvesting*

Q8: What types of content are in your collection? [*for content-holding institutions*; *checkboxes for each content type*]

A simplified categorisation of your core collection areas, to enable an analysis of the diversity of content held under preservation care.

- *Unstructured documents (e.g. ebooks, ejournals, office documents, PDFs, etc.)*
- *Structured documents (e.g. spreadsheets, CSVs, XML, etc.)*
- *Container formats (e.g. zips, tars, gzips, etc.)*
- *Images (2D, still images)*
- *Audio*

- *Video*
- *Database / database records*
- *Scientific datasets*
- *Personal archives*
- *Web archives*
- *Disk images*
- *Software (e.g. applications or games)*
- *Hardware or environments (physical or virtual machines)*
- *Digital artworks*
- *Geographic formats (e.g. GIS)*
- *3D formats (e.g. CAD or point-cloud scans)*
- *Other [please specify]*

Storage capacity

The next two questions ask for an indication of the overall capacity of your digital library or volume of content you hold today, and the expected growth rate over the next 12 months.

This will allow an analysis of storage scales and need for processing infrastructure.

Q9: What is the overall capacity of your digital library? [*for content-holding institutions; checkbox for capacity ranges*]

- *< 1 terabyte*
- *1tb- 10tb*
- *11tb - 100tb*
- *101tb - 1 petabyte*
- *> 1 petabyte*

Q10: By what amount do you expect the capacity of your digital library to increase over the next 12 months? [*for content-holding institutions; checkbox for growth rate*]

- *0%*
- *1-10%*
- *11-25%*
- *26-50%*
- *51-75%*
- *75-100%*
- *Don't know*
- *It will decrease*

About the technology you use (1)

This section asks for information about your production environments, repository system, and workflow tools. The second part asks about your use of services and software tools.

Q11: Production environments [*for content-holding institutions; multi-select checkboxes*]

Please select which of the following infrastructure components are part of your production environment. This will enable analysis of different deployment strategies, for example whether infrastructure is **local** or **outsourced**, delivered through **consortium** arrangements, or through generic **cloud** services.

You can answer for each of these options against the following infrastructure components:

- *Content (e.g. preservation storage)*
- *Metadata systems (e.g. catalogue or discovery)*
- *Processing (e.g. Hadoop)*
- *Repository system (see Q15)*

The options are defined as follows:

- ***Local:** the component is hosted and supported within my organisation.*
- ***Outsourced**: the component is managed by an external company under contract (e.g. Arkivum, DuraCloud, Preservica Hosted, etc.)*
- ***Consortium**: the component is shared amongst other organisations with similar requirements (e.g. DPN, LOCKSS, MetaArchive, etc.)*
- ***Cloud**: the component is provided by a generic cloud company (e.g. Amazon S3 or EC3, Microsoft Azure, etc.)*

If you have multiple arrangements for a particular component, e.g. a local repository system as well as participating in a LOCKSS consortium, you can select multiple checkboxes.

Q12: Operating systems [*for content-holding institutions; multi-select checkboxes*]

What operating systems are used or supported in your digital preservation infrastructure, for servers and processing workstations (e.g. curator desktops)? Servers are hardware that provide services to other users or systems. They may be physical or virtual machines. Processing workstations means specialist machines rather than general purpose PCs. These are often built for a particular task such as creating disk images. This will highlight trends in the use of proprietary or open source operating systems.

- *Windows*
- *Linux*
- *Unix*
- *OSX*

Open-Source

We would like to understand your use of open-source technologies and your level of familiarity with open-source development practices. There are two parts to this. This will allow analysis of the level of engagement with open-source technology.

Q13: Which of the options below best describes your organisation's use of open source technologies? [*mandatory; checkbox for usage*]

Your current use of open-source:

- *We don't use any open-source technology*
- *We use a mixture of open-source and proprietary technology*
- *We use entirely open-source technology*

Q14: How familiar is your organisation with open source development practices? [*mandatory; checkbox for familiarity*]

Your familiarity with open-source development practices:

- *We don't contribute to or maintain open-source technology projects*
- *We contribute to collaborative open-source technology projects*
- *We maintain or lead collaborative open-source technology projects*

Repository Systems and Workflows

The next two questions focus on which systems and workflows are being used in **production** or are under **evaluation**. These questions will help establish trends in the current use of different systems and workflows, and will also identify their existing and anticipated importance.

Q15 and Q16 comprise of two drop-down menus.

Usage

- Production
- Evaluation

(Evaluation might mean you are testing it right now, or it might indicate an anticipated interest in the future. The default is blank - which means 'not in use' or not interested).

Importance

The second drop-down indicates its importance in either of these contexts from 1-5, with 1 being low importance and 5 high importance.

Please refer to the following scales for **production** and **evaluation** in Q15 and Q16:

> **Production:** systems or workflows that are part of operational or production systems providing live business processes. For these items the second drop-down indicates the degree to which your organisation relies upon the system or workflow.
>
> 1. *Used for only one content type or in one workflow.*
> 2. *Used for only a few content types or in a few workflows*
> 3. *Used for some content types or in some workflows.*
> 4. *Used for most content types or in the majority of workflows.*
> 5. *Used for all content types or in all workflows.*
>
> **Evaluation:** systems or workflows that you are testing or using in research projects but which are not used in live business processes. For these items the second drop-down indicates your current or expected level of interest.
>
> 1. *Unlikely to have any interest or requirement in the future, but it's possible.*
> 2. *Might have some interest or requirement in the future, not sure.*
> 3. *Likely to have an interest or requirement, but may not know how important.*
> 4. *Moderate interest or requirement anticipated.*
> 5. *High interest or requirement anticipated.*
>
> **Not in use:** this is the default answer, you do not need to enter this for every piece of software you do not use. If the answer is left blank we will assume not in use.

Q15: What repository system do you use at the core of your digital preservation infrastructure? [*for content-holding institutions; question format described above*]

What repository system do you use at the core of your digital preservation infrastructure (if any, or more than one)?

- *BRICKS*
- *Digital Commons*
- *DigiTool*

- *DSpace*
- *Eprints*
- *Fedora*
- *Fedora (with Hydra)*
- *Fedora (with Islandora)*
- *Greenstone*
- *Invenio*
- *iRods*
- *LOCKSS*
- *Preservica (Safety Deposit Box)*
- *RODA*
- *Rosetta*
- *Bespoke (in-house)*
- *Other [please specify]*

Q16: What workflow tools do you use to manage ingest or content production? [*for content-holding institutions, question format described above*]

What workflow tools do you use to manage ingest or content production, or to coordinate other processing actions? Along with your repository system, this will allow analysis of the need for integrating software products.

- *Archivematica*
- *Goobi*
- *Taverna*
- *Web Curator Toolkit*
- *Bespoke (in-house)*
- *Packaged with repository system*
- *Other [please specify]*

About the technology you use (2)

The second part of this section asks about your use of services and software products.
Both the services and software sections also provide a free text box where you can note comments about the services or software you have identified as important to your organisation.

Brief descriptions of the services and software we have listed, along with links to further information, can be found at the end of this guidance document. Please also tell us about anything which isn't already on the lists - we couldn't include everything!

Usage

Q17 and Q18 also have two drop-down boxes to indicate whether the services and software are being used in **production**, **evaluation** or **not in use**, and a scale from 1-5 in indicate their importance.

Importance

Production: services or software that are part of operational or production systems providing live business processes. For these items the second drop-down indicates the degree to which your organisation relies upon the service or software.

1. *Used for only one content type or in one workflow.*
2. *Used for only a few content types or in a few workflows*
3. *Used for some content types or in some workflows.*
4. *Used for most content types or in the majority of workflows.*
5. *Used for all content types or in all workflows.*

Evaluation: services or software that you are testing or using in research projects but which are not used in live business processes. For these items the second drop-down indicates your current or expected level of interest.

1. *Unlikely to have any interest or requirement in the future, but it's possible.*
2. *Might have some interest or requirement in the future, not sure.*
3. *Likely to have an interest or requirement, but may not know how important.*
4. *Moderate interest or requirement anticipated.*
5. *High interest or requirement anticipated.*

Not in use: this is the default answer, you do not need to enter this for every piece of software you do not use. If the answer is left blank we will assume not in use.

Q17: What third-party services do you use, and how important are they to your digital preservation programme? [*for content-holding institutions; question format described above*]

A list of suggested services is included in the survey and described in the appendix, but please also tell about other services which are important to your institution.

Q18: What software products are you using or evaluating, and how important are they to your digital preservation programme? [*for content-holding institutions; question format described above*]

What software products are you using or evaluating, and how important are they to your digital preservation programme? A list of suggested products is included in the survey and described in the appendix, but please also tell about other software products which are important to your institution.

Q19: In house software [*optional; free-text for name, functional description, comments*]

Please describe any software you are developing in-house - this will help identify functionality which is not currently provided by community solutions.

Q20. Contact details [*optional; free-text*]

If you would like us to email you when the data from the survey are available please provide your name and email address.

Appendix

Services

Service	Description	Links
PRONOM	Online technical registry service hosted by the UK's National Archive.	PRONOM home page
TOTEM	Trustworthy Online Technical Environment Metadata registry.	TOTEM home page
Plato	Web based preservation planning service, Plato is a decision support tool that guides the user through the preservation planning process.	Plato site Plato GitHub project
Scout	A web based preservation watch system. Scout provides an ontological knowledge base to centralise information to detect preservation risks and opportunities.	Scout GitHub site
bwFLA	Browser based, on demand access to emulated software environments.	bwFLA web site
COPTR	Community owned digital preservation tool registry, collates the digital preservation community's software tool knowledge.	COPTR home page
Q&A	Digital preservation and question and answer site, where anyone can ask DP related questions to be answered by other members of the community.	Q&A site home page
OPF KB	The Open Planets Foundation wiki, a shared workspace where people can collaborate to produced online documentation.	OPF wiki home

Digital Preservation Tools

Production Tools

Tool Name	Description	Links
DROID (Digital Record Object Identification)	Format identification tool.	TNA DROID page DROID GitHub project
FIDO	Format identification tool.	OPF FIDO page FIDO GitHub project
*nix File utility	Format identification tool packaged with linux / unix distributions.	The Fine Free File Command
Apache Tika	Format identification / Characterisation / Content extraction	Apache Tika project Apache Tika SVN Tika GitHub project
JHOVE (JSTOR/Harvard Object Validation Environment)	Format identification / Format validation / Characterisation	JHOVE Site JHOVE GitHub project
JHOVE 2	Format identification / Format validation / Characterisation	JHOVE 2 Wiki
FITS	Format identification / Characterisation	Harvard FITS site FITS GitHub project
NZ Metadata Extractor	Characterisation	Metadata Extraction on SourceForge
Jpylyzer	JP2K Validation software	jpylyzer Jpylyzer GitHub project
SIARD	Relational database archiving software	Swiss Federal Archives SIARD page
DD Rescue	Data recovery / disk imaging tool	GNU ddrescue page
ExifTool	Characterisation and metadata manipulation tool	ExifTool web site
MediaInfo	Unified display of the most	https://mediaarea.net/en/Med

	relevant technical and tag data for video and audio files	iaInfo
FFMpeg	Video Characterisation and Migration software	FFmpeg home page
Heritrix	Web crawling/archiving software	Heritrix Wiki
NetarchiveSuite	Plan, schedule and run web harvests of parts of the Internet	NetarchiveSuite site
ImageMagick	Image toolbox	ImageMagick site
Bagit	File packaging and transfer format for arbitary digital content,	BagIt draft specification Library of Congress transfer tools
MIXED	Structured data migration to XML	MIXED site
ODF Validator	Open Document Format validator	Apache ODF Validator site
EpubCheck	Validates EPUB files	EpubCheck site
TrID	Format Identification	TrID website
BitCurator	Digital forensic tools for collecting professionals	BitCurator site
PDFBox	Command line tools and Java library for working with PDF files.	Apache PDFBox home page

Prototype Tools

Tool Name	Description	URL
(from SCAPE)		
Flint	PDF/EPUB policy validation tool	Flint site
xcorrsound	Audio waveform comparison toolbox	xcorrsound site
matchbox	Near-duplicate image detection	Matchbox site
Crop Detection Tools	Detect image cropping errors in scans of text	Crop Detection Tool GitHub project
pagelyzer	Assess the similarity of web pages for web archive quality control	Pagelyzer site
Nanite	Scalable identification tool based on DROID	Nanite site
ToMaR	Execute command line tools within Hadoop	ToMaR site
HAWARP	Hadoop based web archive processing	HAWARP site
C3PO	Collection profiling tool	C3PO GitHub site
(from TIMBUS)		
DP Expert Suite	Virtualisation manager for archiving and accessing legacy software systems.	Digital Preservation Expert Suite Demo
Extractor platform	Software to extract features of a hardware and software environment as preservation metadata.	Context Model Extraction Framework Demo
Legality Lifecycle Management Tool	Decision guidance tool that helps the user establish digital preservation legal contexts.	Legality Lifecycle Management Tool Demo