This survey was undertaken by the Archives and Records Council Wales to provide an evidence base for developing a national digital preservation service for ARCW members.

ARCW Digital Preservation Survey Report

Jake Henry, ARCW Project Manager, February 2017

Contents

Executive Summary

This Digital Preservation Survey Report demonstrates that the need to store, manage and provide access to digital information is a real and immediate issue. The lack of a national preservation infrastructure poses a real threat to current governance and accountability and risks significant gaps in the documentary heritage. A national preservation solution which includes a technical infrastructure, supported by the use of standards, workflows and documentation is essential to enable continued access to our national heritage.

The ARCW Digital Preservation Working Group, which has been working collaboratively since 2008 to increase capacity, has just produced a digital preservation policy for Wales. This policy states that archive institutions have a vital role to play in safeguarding the documentary heritage. This heritage is becoming increasingly digital, but in the present financial climate, there are many and competing demands on financial and staff resources, which means a collaborative approach is the only feasible option.

This collaborative approach fits with the strategic vision laid out in the Archives for the 21st Century national strategy which recommended a 'co-ordinated response to the growing challenge of managing digital information so that it is accessible now and remains discoverable in the future' [1], and remains a strategic priority for the sector going forward.

This pressure on archive services to have a long-term digital preservation solution at their disposal is further reinforced by the need to meet or continue to meet the Archives Service Accreditation standard.

This survey provides an evidence base to support the development of a national digital preservation solution. Any solution has to be sustainable, scalable and open to integration/interaction with other systems. The survey confirms that the needs of the archive services, research community and records management sectors are similar in that secure, scalable, long-term storage is required. However, the complexities which arise from records deriving from different sectors – for example:format, retention periods, size, legal frameworks, cross organisation department access demands and the complex nature of all the different stakeholders – it may be more appropriate to store some classes of data locally or to develop ingest workflows from current database systems into a national digital preservation system rather than developing one system for all.

For records of permanent historical significance transferred to Archives one approach which has already been developed for ARCW members links Archivematica, an open source solution which undertakes preservation actions, to Fedora (another open source solution), which manages the digital content. This approach enables digital content to be deposited in the system through a web interface, where it can be managed and preserved, with the aim of providing access through links to other systems, which may be national or local.

Solutions are being developed and skills and knowledge are increasing. The proposed technical solution, which could be developed as "A Trusted Cloud for Wales", which is supported by the

[1] Archives for the 21st Century (2009), p.16.

national preservation policy and institutional strategies, will ensure that Wales's documentary heritage remains authentic, trustworthy and accessible now and in the future. This solution will provide archive services with the ability to meet national strategic priorities and Archives Accreditation. This is achievable, providing that development costs can be secured and strategies for managing the costs for sustaining systems in the longer term are put in place.

ARCW / IRMS Wales Digital Preservation Survey Report

Context

In 2009, Archives and Records Council Wales (ARCW) established a Digital Preservation Working Group (DPWG) to address specifically the requirements for the preservation and management of digital records. The establishment of the DPWG recognised the particular challenges associated with the preservation of digital material, notably the fast pace of software and hardware development, the increasing complexity of digital resources, and the resulting impact on the stability of such media. If digital material is to remain accessible, both in the short-term (for business continuity, research and economic / legal requirements), and for longer-term cultural preservation, then measures have to be taken at an early stage, otherwise there is a huge risk that the information will be inaccessible or lost.

One of the first actions of the group was to undertake a survey of digital preservation in Wales to obtain a "state of the nation" picture of preservation activity, to raise the profile of digital preservation, to highlight knowledge gaps, to identify training needs and provide evidence for applying for funding streams. The survey identified that barriers to digital preservation included lack of technical infrastructure and storage, skills, resourcing and knowledge. Following the survey, a business case was submitted to the Welsh Government via its Museums and Archives and Libraries Division and funding was secured to employ a project officer. A practical working group was also established to pilot differing solutions and to develop capacity, skills, workflows and documentation.

Aim of the survey

The aim of this second survey was to provide evidence of the current state of the digital preservation landscape in Wales. This information would then be used to inform the development of national digital preservation architecture. The survey sought to gain information about potential users and use cases, storage capacity, any existing digital preservation activities and training requirements.

This report is based around the information gathered through the survey which was produced as an online questionnaire titled *"ARCW / IRMS Wales Digital Preservation Survey, Arolwg Cadwedigaeth Ddigidol ARCW / IRMS Cymru"* which was made available on the National Library of Wales website, initially between 23/02/2016 and 05/04/2016 and then again between 29/04/2016 and 06/06/2016 after requests to extend the contribution period. In total we had 16 responses.

The new survey was based on multiple choice questions with room for extra information through comments boxes providing both quantitative data about the areas specifically of interest to the ARCW Digital Preservation Group and the opportunity for qualitative comment. The aim was to receive responses from as broad a group of people as possible, including local authority archives and records management services and educational bodies.

Questions Asked

Question 1 - In which sector do you work?

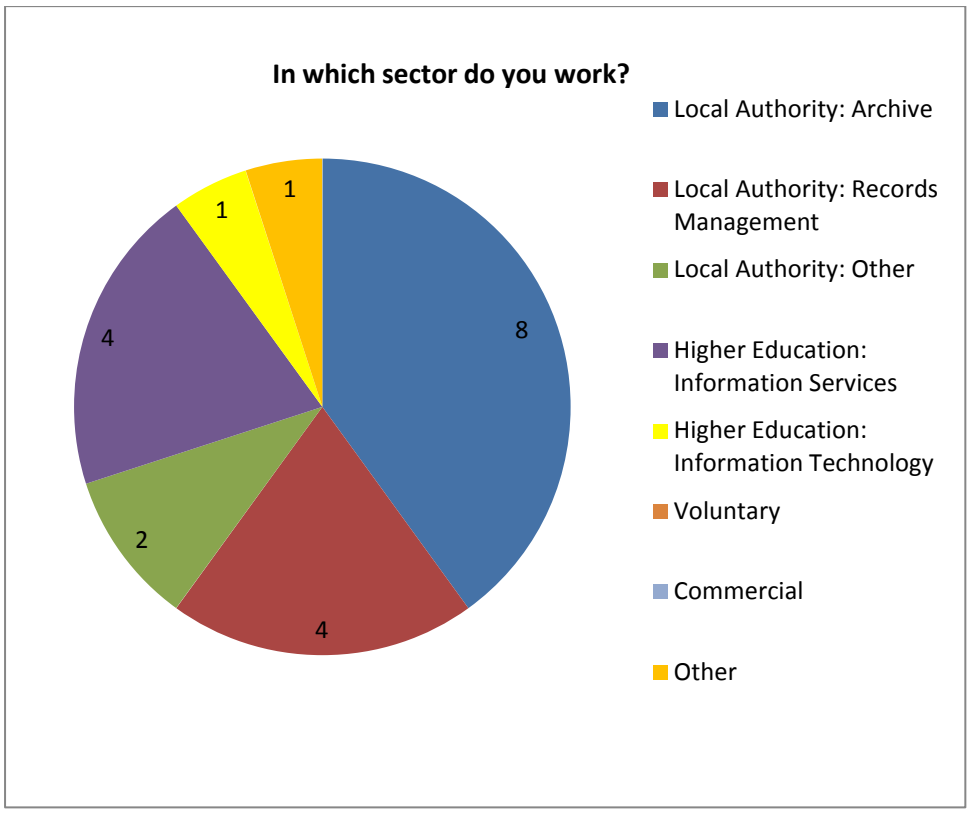

Figure 1

As the majority of the ARCW Digital Preservation Group is made up of members representing local authority archives services, it is not surprising to see that they also make up the majority of responses to the survey. Multiple responses also received from representatives of the education and records management sectors..

Question 2 - Do you have digital information which needs to be accessed and retained in a managed and secure way?

.

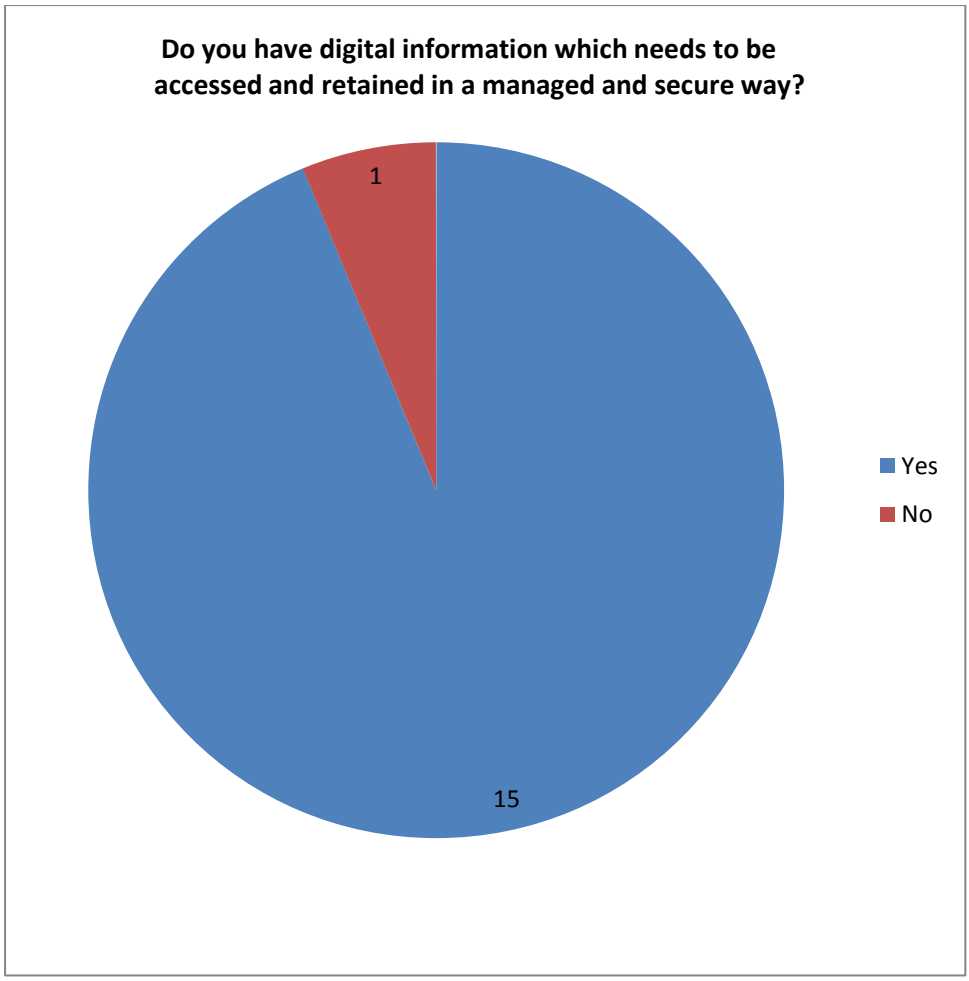

Figure 2

The vast majority of responses confirmed that secure storage and management of digital data is a real and immediate issue.

Question 3 - How long does the digital information need to be retained?

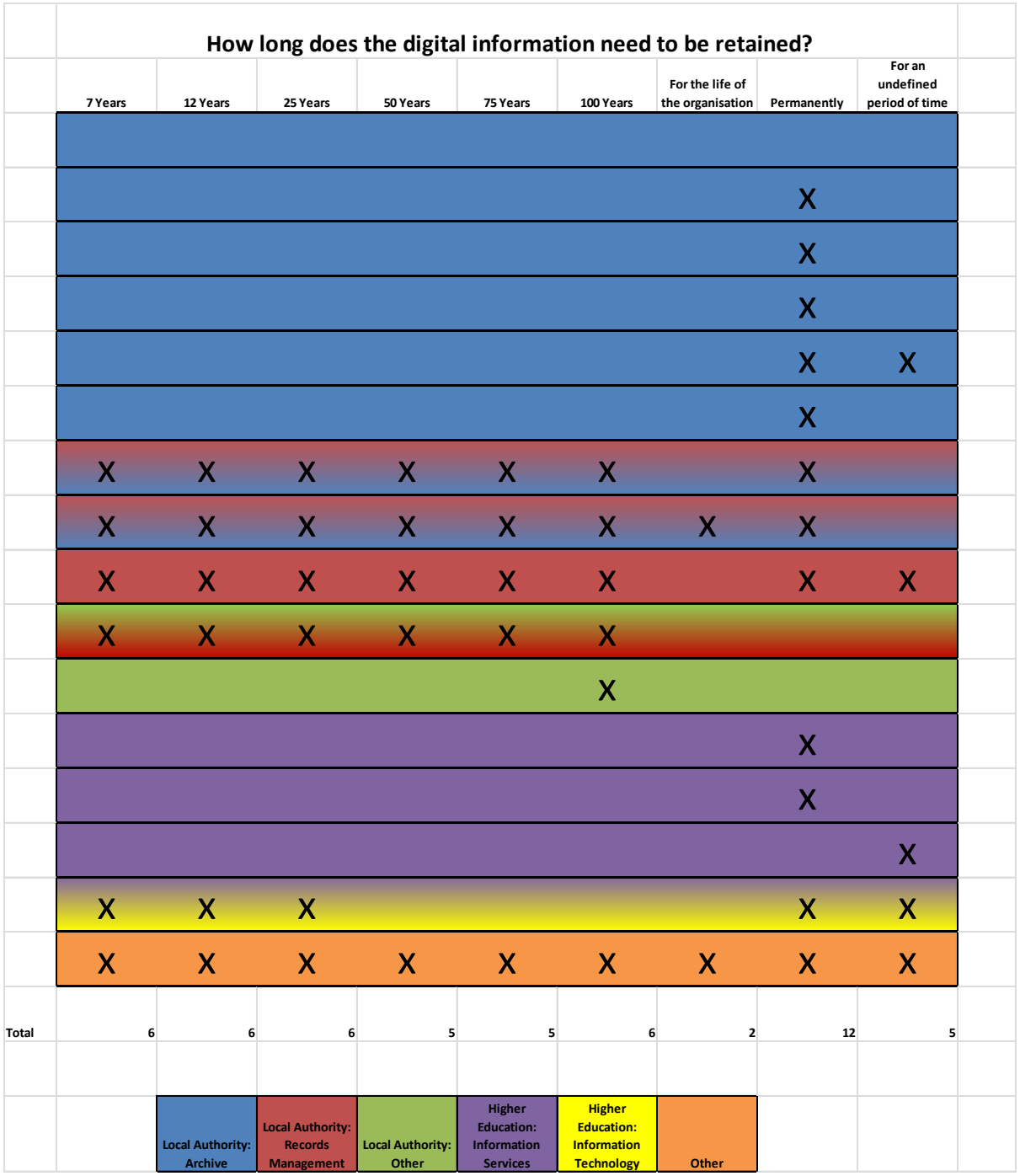

Figure 3

As may be seen from the chart above, respondents from a local authority archive or higher education background all have digital information which either needs to be stored permanently or for an unspecified period of time only. This is supported by comments such as:

> *"Born digital archives, and digital surrogates of archives and rare books, are for permanent preservation. The University's institutional repository has a commitment to retain items indefinitely."*

And

> *"All material that we are responsible for is designated for permanent retention."*

However responses from other areas, especially local authority records management, indicate that there is much more variety in the lengths of time information needs to be kept.

> *"If including current and semi-current Council records, plus records deemed of permanent value for the Record Office we would cover the entire spectrum of retention periods."*

Question 4 - What is the purpose of storing the digital information?

The purpose for storing information could affect the requirements for storage provision. For example, items needed for continued business use would need to be easily accessible whilst those stored for historical preservation may not need to be accessed on a regular basis. Often a combination of uses would be needed. For example, an archival copy could be stored for historical preservation on tape, whilst a derivative access copy would be made available online to sustain public access.

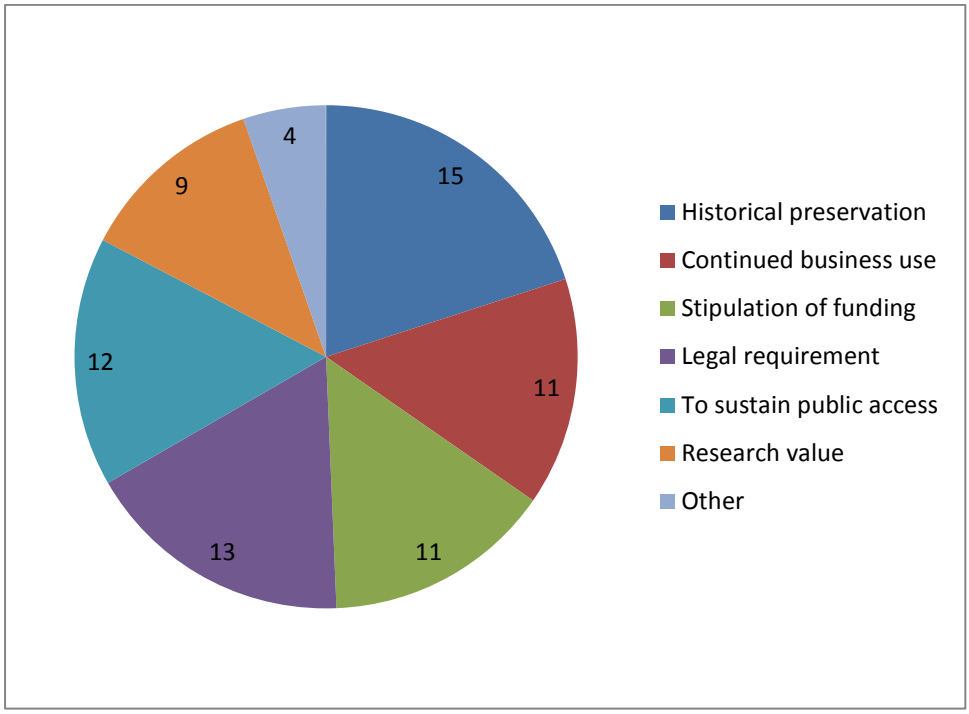

Figure 4

Looking at the chart on the previous page, and considering that 16 individual responses were received to the survey, it is apparent that the majority of respondents require the storage of data for multiple purposes. This highlights the question of whether this is best provided as a single service, or multiple services tailored to achieve the different requirements.

Question 5 - What types of records are you most likely to be storing in a Digital Preservation System?

The responses to this question were so broad that it is almost impossible make a meaningful analysis. The content stored by respondents from an archive service appears to be more varied than those from a records management. However the responses to this question are best summarised by two quotes:

> Question
> *"What types of records are you most likely to be storing in a Digital Preservation System?"*
>
> Answer (Local Authority Archive)
> *"Those that need to be retained longer term"*
>
> Answer (Local Authority Records Management)
> *"Any record currently held electronically"*

Question 6 - Which file formats do you hold and will need to be retained in the planned digital preservation system?

Following on from the previous question, the responses confirm that the type of files in need of preservation is very varied. Even the categories which had the least responses ('Vector Graphics' and 'Other') are still marked as formats which need to be retained by approximately 30% of people or respondents.

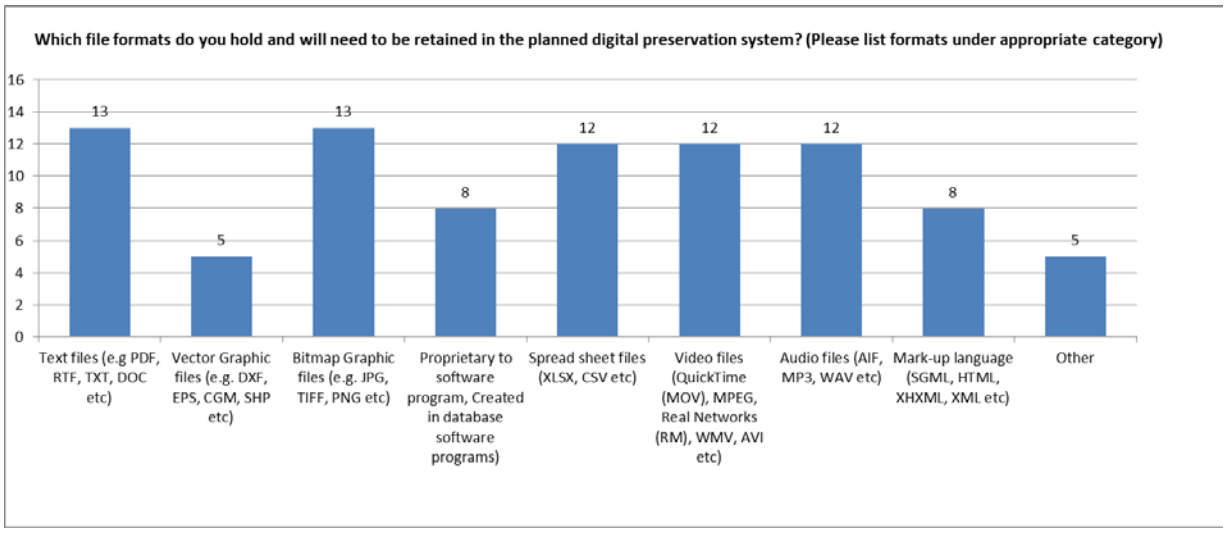

Figure 5

When considering the responses provided in the comments boxes, it is clear that this is an area still very much in development and people are not fully aware of the files they may need to store in the future:

"Ddim yn ymwybodol eto o fformatiau gwahanol sydd angen eu cadw" ([I am] Not aware as yet of other file formats [we have] that need to be preserved)

There is however awareness that there will be a need to preserve many different types of files and formats:

"I think it is safe to say that many of the above would be included."

And

"There are probably formats other than those listed but the above should give a flavour."

Question 7 - What kind of information (metadata) is held about the records?

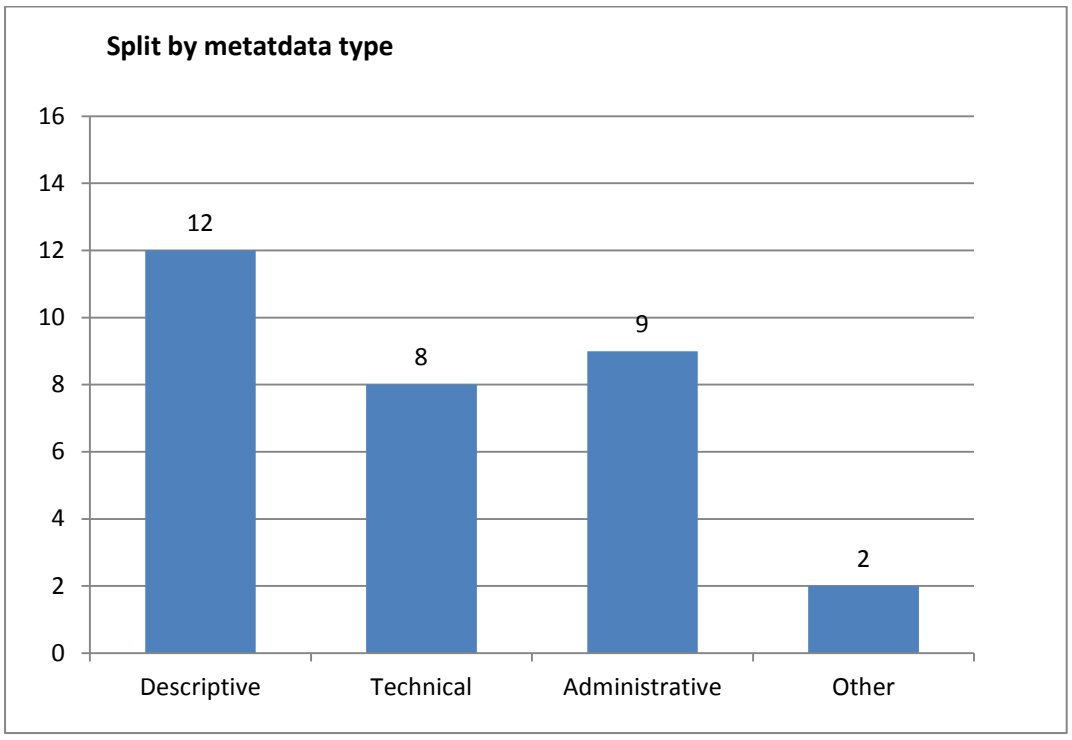

Figure 6

Splitting the responses up by metadata type indicates that metadata of all types is being stored by many of the respondents. Descriptive metadata is most commonly stored, with three quarters of respondents storing this type of metadata. Technical and/ or administrative metadata is also stored by half of the people taking part in this survey.

If the data is displayed differently, splitting it instead by respondents (as shown in Figure 7 on the following page) it is clear that nearly a third of those surveyed are storing descriptive, technical and administrative metadata. This is very encouraging, as it shows a strong understanding of the need to store metadata and will mean that appropriate metadata is available for any future digital preservation solutions.

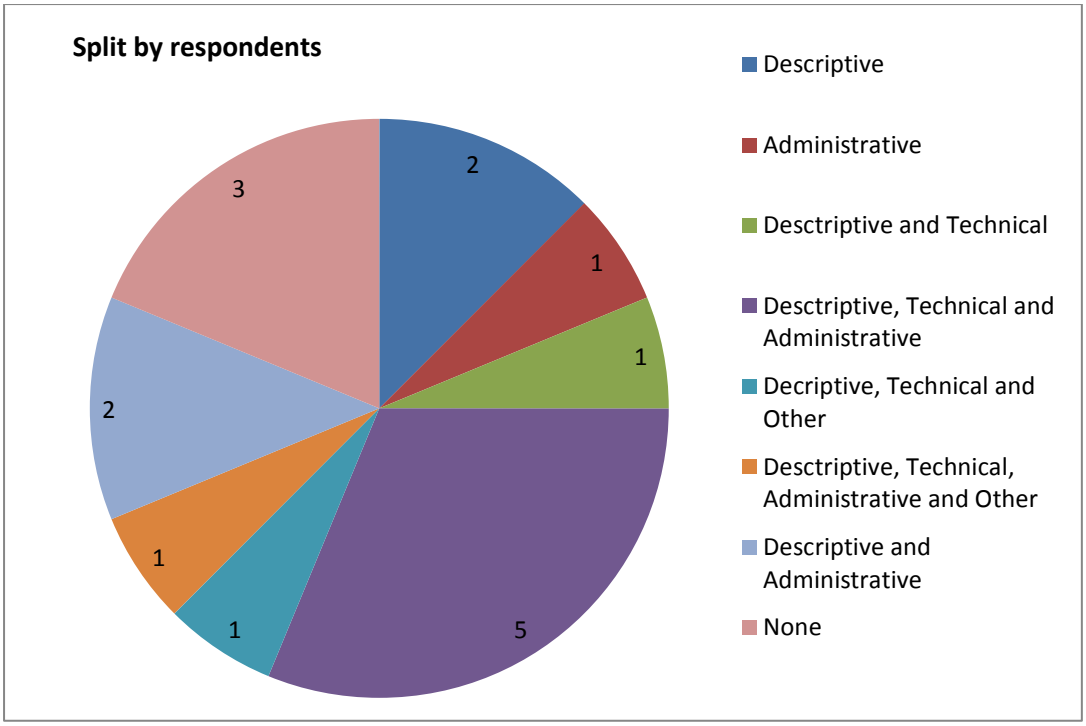

Figure 7

Of more concern are the three survey respondents who did not report holding any metadata about their records. When considering the details provided by these respondents, it is clear that they do recognise the importance of storing metadata, but they either felt unable to give an accurate answer (due to the amount and variety of records stored and possible differences in the way metadata was collected and stored for different types of files) or were at the early stages of developing digital preservation workflows and had not yet developed a system for collecting and storing metadata.

Question 8 - What existing systems are in place for the control and management of digital information?

The consideration of existing systems is important as it provides an indication of how much infrastructure is already in place to preserve digital content. It also supports the information gained from the previous question, giving a better understanding of the kinds of metadata being held. This information also provides an indication of what type of systems any future digital preservation solution will need to integrate with (although this is addressed more thoroughly in the next question).

It is apparent that over half of those taking part in the survey are either using some archival cataloguing system (e.g. CALM), an in house solution, or some combination of the two. EDMS and EDRMS are being used less; however there is still a significant amount of use. In terms of specific packages, SharePoint is currently being used by four of the respondents and is under consideration by another.

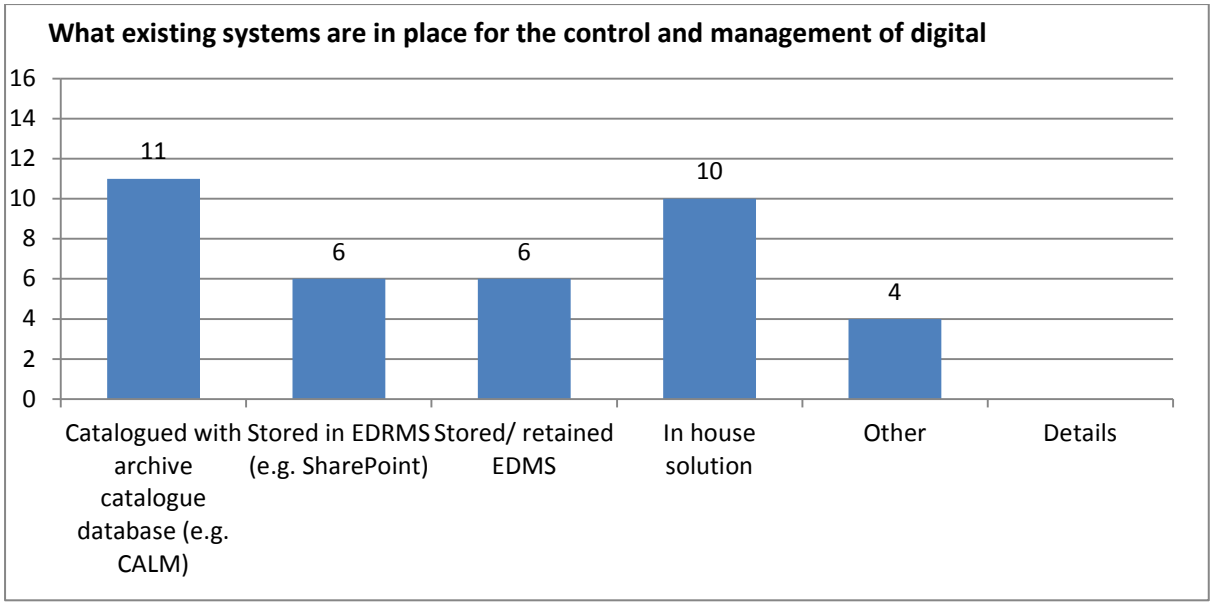

Figure 8

Question 9 - Which other systems need to integrate with the digital preservation solution?

Numerous software packages were mentioned, although some software packages are commonly used. This is confirmed by the fact that CALM is mentioned by seven of the 16 respondents. In total 16 different specific pieces of software were mentioned with 14 of them only being mentioned by a single institution. This suggests that it may be sensible to develop a generic system for integration of many different pieces of software, which would enable easy integration between different systems and require minimal work to customise to any specific software package.

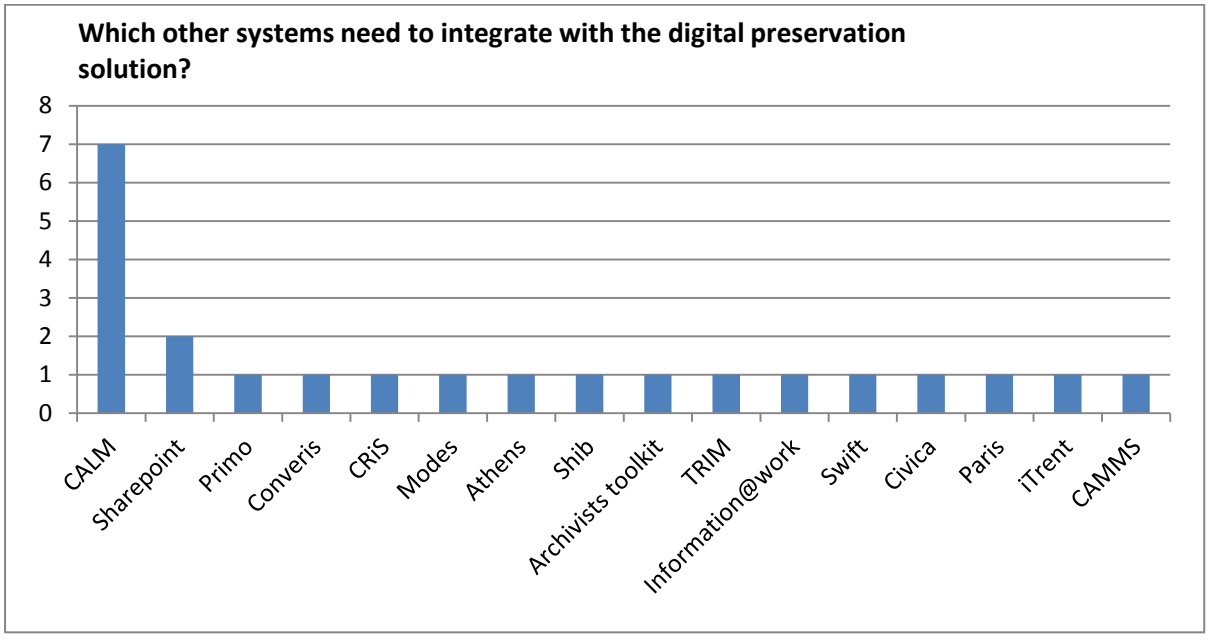

Figure 9

Storage

The consideration of the amount of storage which will be required is one of the most basic technical issues to consider when building a digital preservation system. It is also inherently one of the most difficult questions to answer accurately. The answers to question 3 indicate that the majority of respondents have at least some data which needs to be stored permanently. This data will grow and the amount of storage needed will increase over time, as more content is acquired and created, while much of the existing content will not be being removed.

It should also be stressed that any figures, however accurate, gathered on the size of the digital information in its current format, will not directly equate to the amount of space needed to store it securely. Secure archiving of digital content will require not only the original file to be held, but also a 'normalised' preservation copy (which in many cases will be far larger than the original[2])

This survey has aimed to provide an accurate a picture of the storage needs, both now and in the future, by splitting the question into two parts. The first part asks how much data is currently held which needs secure storage, and the second part asks how much more storage may be required over a 12 month period. With the information provided by these questions and combining this information with the expected file types listed in question 6, it will be possible to extrapolate a figure to predict how much storage may be needed in the future. It should be noted that as born-digital becomes the default means of recordkeeping storage requirements are only likely to increase.

Question 10 - Approximately how much digital information needs to be stored securely?

As is apparent from Figure 10 on the next page, there is a neat split between those who have 100 – 999 GB and those with over a TB of digital information in need of secure storage. Only one respondent currently has under 100GB, although they did comment that:

> *"there may be more that has not been identified at present."*

Three respondents did not answer the questions, but provided comments indicating that they did not know, or did not have access to this information. This could be because the information is held by a different individual or department, but is also possibly due to the inherent difficulties of finding this kind of technical information, without systems which provide the information easily. In some cases, the digital information which requires secure storage will be spread across multiple locations and storage types. With information held on a variety of servers, optical discs, local computers and external hard drives, this is a clear example of the need for digital preservation policies, strategies and workflows.

[2] An original JPEG image from a 10 megapixel digital camera could easily be under 1 MB. However the preservation TIFF file would be slightly over 32MB. This is due to compression which poses an innate preservation risk.

The vast majority of organisations who responded hold over 100GB of information needing secure storage. The comments provided indicate that some have vastly more than 1TB. The highest figure mentioned being 150 TB, but there does not appear to be any specific correlation between the kind of organisation and the amount of digital information to be stored.

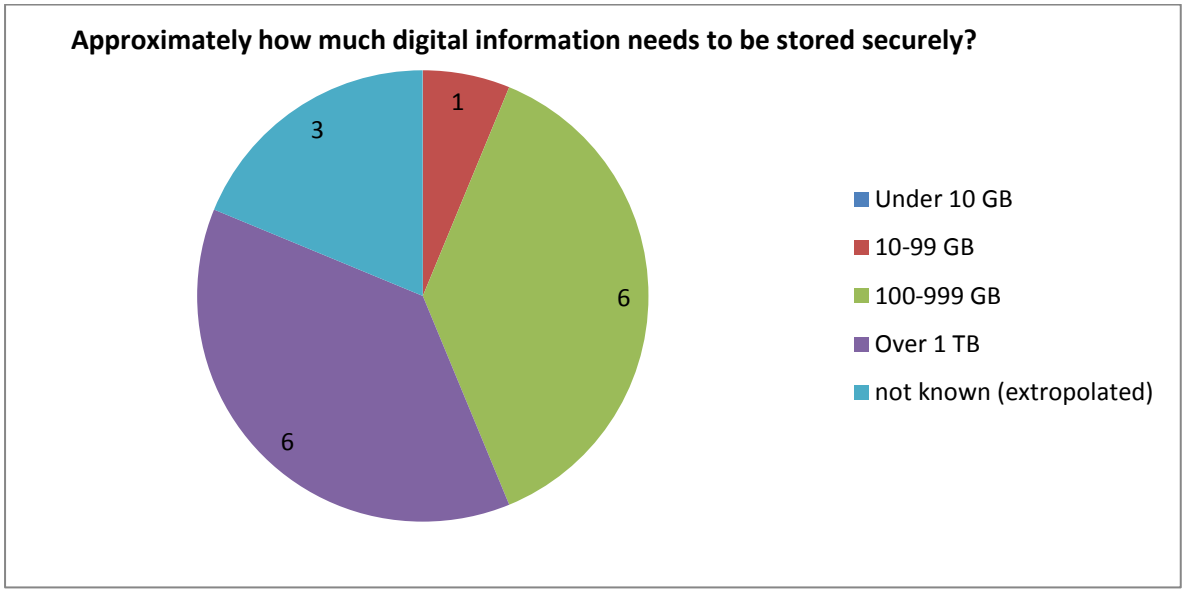

Figure 10

Question 11 - Approximately how much digital information would you expect to produce/ receive over a 12 month period?

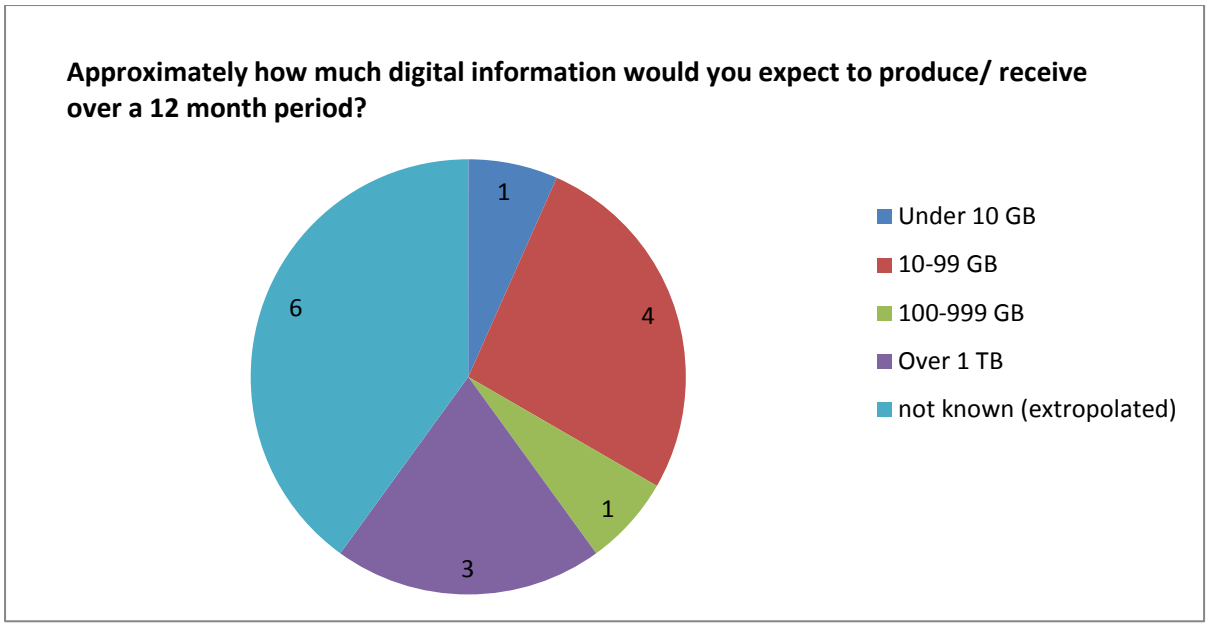

Figure 11

Figure 11 above shows that the majority of respondents did not answer the question. It is assumed that they did not respond as the information was not available[3]. This highlights the inherent difficulty in predicting this sort of information. The comments indicate that the nature of the information makes it difficult for accurate estimates of the amount of storage space required.

> *"Due to the varied nature of the material (moving image, stills, OCR, DataSets, Administrative, etc) the quantities are potentially very large, but equally difficult to predict"*

This makes it very difficult to estimate basic technical information, such as the size of the data.

Respondents who answered the question can be broadly split into two categories: those who expect a relatively small amount of data (10 – 99GB) and those who are expecting significantly more (over 1 TB)

Question 12 - Do you currently undertake any digital preservation activities or have a defined workflow for dealing with digital information?

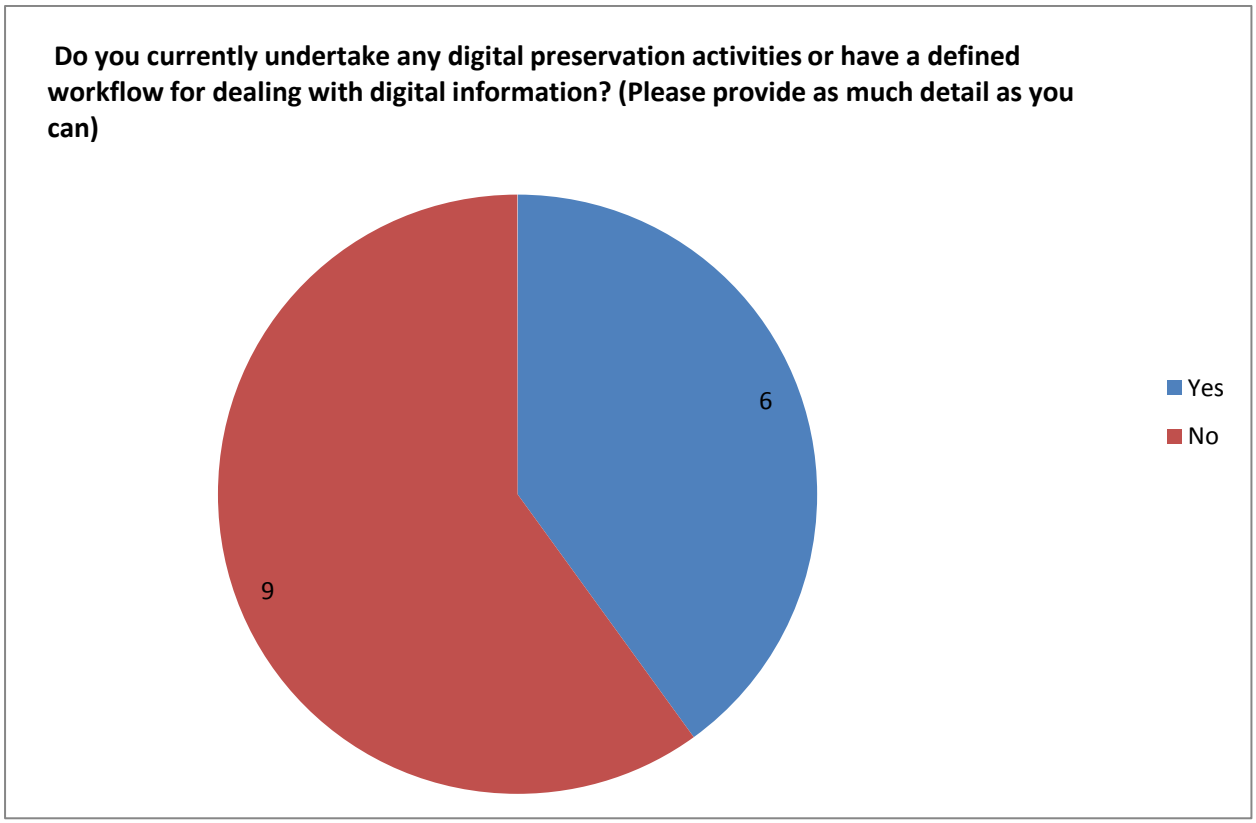

Figure 12

60% of respondents do not currently undertake any digital preservation activities.

[3] This is assumption is backed up by the comments, e.g. *'Ddim yngwybod eto'* and *'Unknown'*

Comments included:

> *"Needed ASAP. Being worked on by our staff"*

and

> *"Currently sourcing equipment and developing workflow to accept born-digital archives"*

Many simply left the comments field blank, which may suggest that the need for digital preservation has been identified[4] but plans to address this need have not yet been developed.

When analysing the comments provided by those who do currently undertake digital preservation activities, the emphasis seems to focus on the creation of backup copies of the information and storing it separately in a centralised location. Apart from this, digital data appears to be treated in a similar way to other, non-digital information.

> *"At present we are undertaking Digital Preservation in a way that mirrors analogue accessioning"*

There was only one mention of format migration, but it is apparent that the issues surrounding digital preservation are being considered.

Question 13 - Do you have a Digital Preservation Policy and/or guidance in place?

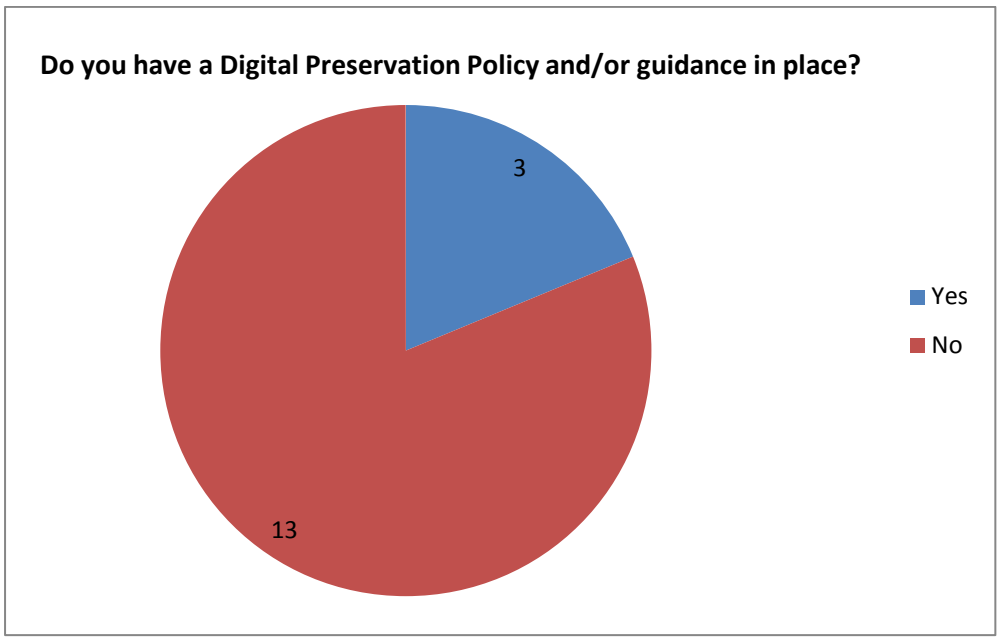

Figure 13

[4] See Question 2 - Do you have digital information which needs to be accessed and retained in a managed and secure way?

These results reiterate the conclusion made from the question above. Of the 3 who answered 'yes' the comments fields read:

"Guidance only, no policy",

"Basic policy"

"We have guidance for processing digital records and a policy in draft form."

This indicates that there is a growing understanding of the issues involved with digital preservation and several comments refer to the intent to create some form of digital preservation policy/ guidance.

Question 14 - Have any negative issues arisen with regards to preservation of digital information?

When considering the survey responses, it may be falsely reassuring to see that only 7 respondents have experienced negative issues with regard to digital preservation. This may have more to do with the limited amount of work being done, rather than positive experiences. If the results are compared with Question 12 - Do you currently undertake any digital preservation activities or have a defined workflow for dealing with digital information? It is apparent that of the 6 people who answered 'yes', 5 also answered 'yes' to having had negative issues. It is also interesting to note that the 2 respondents who answered no to question 12 were still aware of negative issues which had arisen concerning digital information, perhaps suggesting that they do some basic analysis of digital preservation issues, even though they may not have considered these to be digital preservation actions.

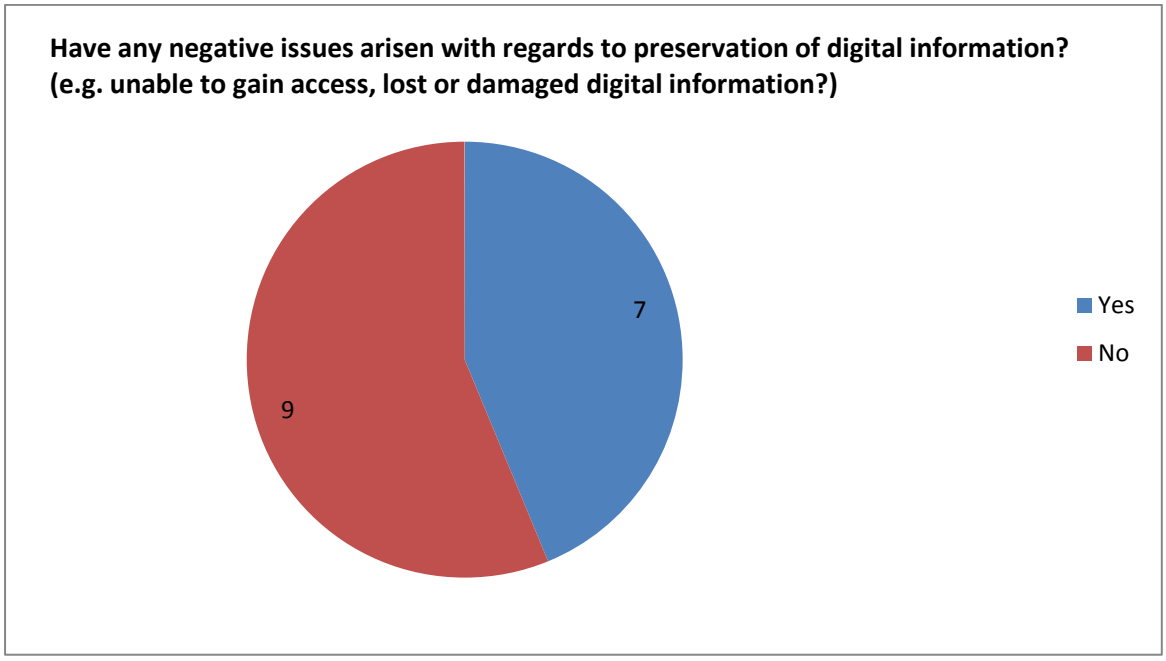

Figure 14

The main issues raised in the comments of those who answered yes are inability to read data from magnetic and optical discs and concerns surround file format obsolescence and propriety formats. There was also mention of issues digitising audio tape. This brings up the question of exactly what respondents mean by digital preservation. The issue of digitising analogue material is not in itself a digital preservation issue. Rather it is a traditional conservation problem until such point as it is successfully digitised. There is perhaps a conversation which needs to be had to ensure everyone understands digital preservation as the same thing.

Question 15 - If guidelines or training opportunities in preservation of digital content were provided which subjects would you find useful?

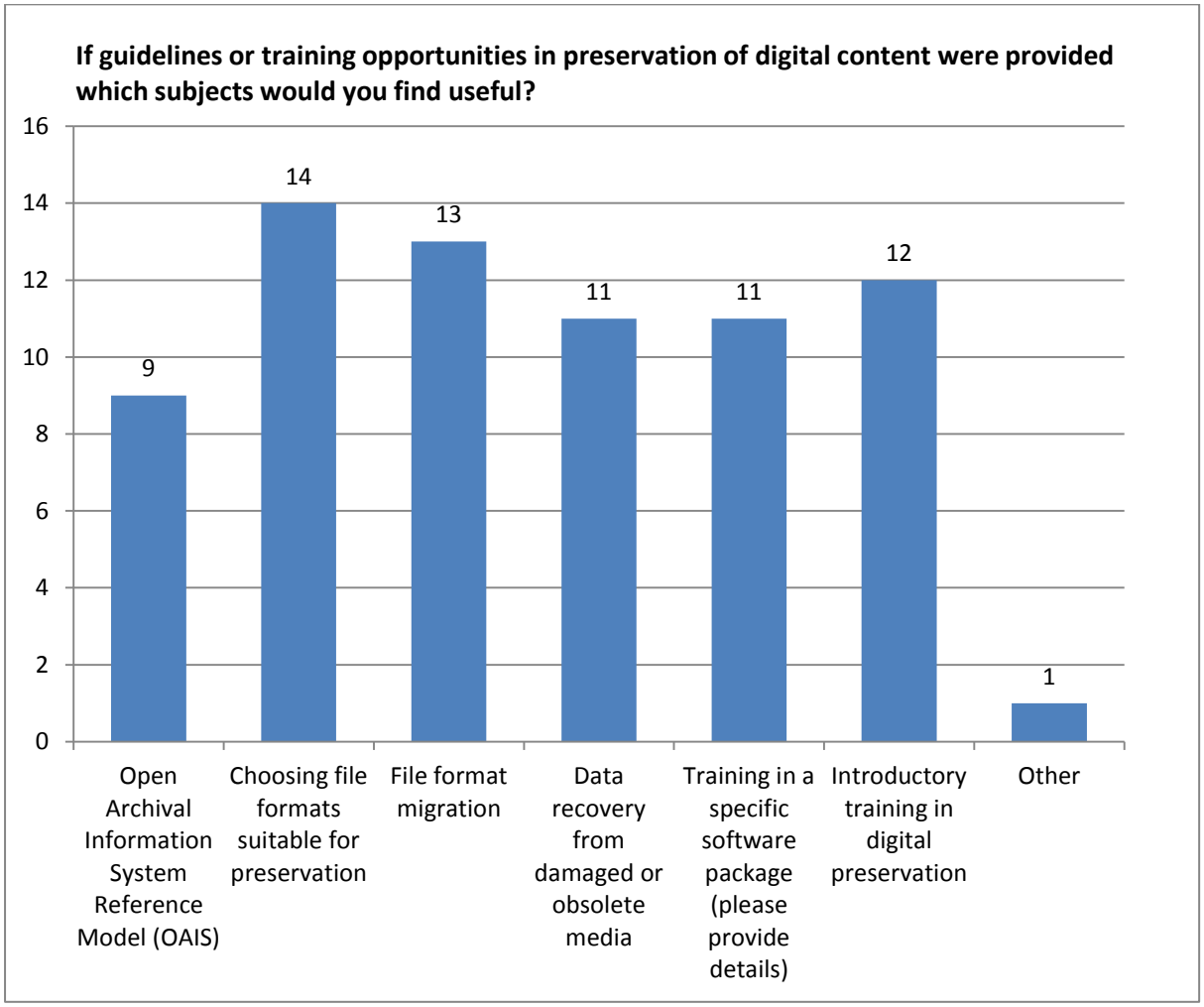

Figure 1513

Comments stated that whilst some members of staff in organisations had taken part in digital preservation training, others had not had the opportunity. It was also suggested that if a digital preservation workflow was created, specific training in this would be required. The one request for other training commented:

"Business cases would be really useful, to see how other organisations have tackled the issues surrounding DP. One of the main issues is that DP spans multiple departments and in larger organisations it can be difficult to know where responsibility sits."

.

Question 16 - How you would like training delivered?

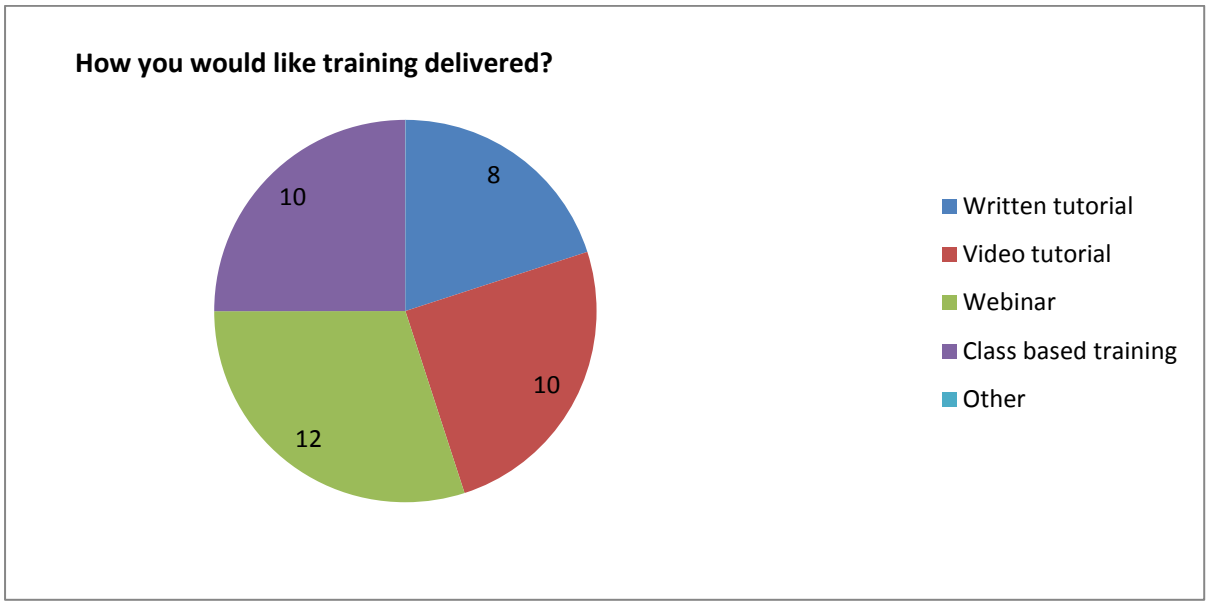

Figure 16

The comments box responses can be summed up well in one quote:

"Each option has its benefits"

Key points

- The majority of the respondents were from local authority archive services
- The need to store, manage and provide access to digital information is a real and immediate issue
- Archive services need to preserve the majority of content permanently
- Records management requirements have differing retention and access needs
- A variety of formats are being created and stored
- Metadata is not being created or managed by all respondents
- Need to integrate/interact with widely used systems such as CALM and Sharepoint,, but with the ability to interact with other systems
- Digital content is being held on multiple locations and storage types, risking duplication of data, lack of intellectual and physical control
- Digital preservation requires lots of storage space, as preservation and access copies are required
- Very few organisations have digital preservation policies in place, but there is a growing awareness of the need for digital preservation

- Training should be delivered in a variety of differing formats, including webinars

Proposed solution

The national digital preservation policy provides the framework for ensuring that digital resources remain authentic and accessible in the future by countering the threats of technological obsolescence and the fragility of digital media. It seeks to facilitate the appropriate management of records which merit permanent preservation, regardless of format. The policy acknowledges the need for active preservation of digital resources, as passively storing digital information on hard drives or removable storage media will not ensure their permanent preservation, or enable continued access.

The solution proposed by the ARCW group is a linkage of the two open source systems, Archivematica and Fedora. These systems would work together to provide all the attributes necessary for the management and preservation of the digital holdings. . Archivematica would be used as the way to get or "ingest" content into the system and Fedora as the preservation solution. Archivematica would provide the ability to ingest authentic, trustworthy and reliable born digital material through its generation of structural information, through METS and record verification, and through its generation of technical metadata, using Droid, JHOVE and virus checking. Fedora would then preserve the content and manage it within a Digital Archive.

This solution was proposed for the following reasons:

- Archivematica was originally selected in 2009 and has been tested by ARCW members. It is open source and does not require much configuration. It provides all the required preservation functionality and integrates with the ATOM cataloguing system, which is used by some ARCW members
- Fedora is a digital asset system which has been used by the Library for over 15 years to manage its digital content. The Library has experience and confidence in its use and performance
- It matches the needs identified in the report, as it enables active preservation, it is scalable, creates appropriate metadata and is suitable for different types of content
- It integrates with other solutions – harvesting of websites into Archivematica from moderngov has been achieved. Generic workflows can be developed to integrate with other systems, including SharePoint and with cataloguing and retrieval systems, including CALM
- It conforms with national standards and has a growing community of use. The Archivematica UK group has representatives from archives, universities, museums and galleries
- Content would be hosted in a NLW cloud, which will provide a secure system, with the management and governance undertaken centrally. This will provide a consistent and cost effective approach for the preservation of data for permanent preservation, but does not prevent the use of local systems for the storage of other categories of data

Conclusion

The survey confirms that the needs of the archive services, research community and records management sectors are similar in that they all require the preservation of trustworthy records in secure, scalable and long-term storage. The proposed Archivematica/Fedora solution is appropriate for records which have been identified for permanent preservation by archival institutions. However, due to the complexities which arise from records deriving from different sectors, for example; format, retention periods, size, legal frameworks, cross organisation department access demands and the complex nature of all the different stakeholders, it may be more appropriate to hold some data on local systems or to develop ingest workflows from current database systems into the national digital preservation system for historical preservation rather than developing one system for all.

The ARCW group has just produced a digital preservation policy for Wales, which states that archive institutions have a vital role to play in safeguarding the documentary heritage. This heritage is becoming increasingly digital, but in the present financial climate, there are many and competing demands on financial and staff resources. Lack of progress in managing, storing and providing access to digital information poses a real threat to governance and accountability now and could lead to significant gaps in the documentary heritage of the future. The proposed technical solution, supported by the national preservation policy and institutional strategies, will ensure that Wales's documentary heritage remains authentic, trustworthy and accessible in the future, but this depends upon the securing of appropriate funding to further develop the solution and for sustaining the service in the future.

.