Open Preservation Foundation

Digital Preservation Community Survey 2015

DRAFT: Initial Findings 0.1

Table of contents

Authors
❏ Carl Wilson, Becky McGuinness, Ed Fay, **Open Preservation Foundation** ❏ Nick Krabbenhoeft, **CodedCulture**

1. Introduction

We ran our digital preservation community survey in December 2014 - January 2015. The goal of the survey was to assess the current state-of-the-art in digital preservation and identify adoption of digital preservation tools and approaches. It followed on from our member survey carried out in October 2014 to build an evidence base about our members, and steer our annual plan. Selected information from our members responses have also been included in the results.

Acknowledgements

Our thanks go to Andrea Goethals, Harvard Library, and Trevor Owens, Institute of Museum and Library Services, for their role in reviewing the survey content and structure, and their support in disseminating it to the community.

Our thanks also go to Nick Krabbenhoeft, CodedCulture, for his additional analysis and contribution to the IPRES 2015 poster.

References

We compared some of our results to the results of a survey carried out by the EU PLANETS project (2006-2010) to assess organisations' preparation for digital preservation to see how today's reality reflects expectations: http://www.planets-project.eu/docs/reports/planets-survey-analysis-report-dt11-d1.pdf (published in 2009)

1.1 Methodology

The responses were collected through SurveyMonkey, and the survey was open for just over eight weeks. We provided additional guidance notes to explain the purpose and format of each question. Many of the questions were optional to allow for responses from different organisation types.

1.2 Response rate

We received 132 responses to the survey.

2. Organisation type and location

The highest response came from libraries; 34% of responses were from academic/research libraries and 12% from national libraries. 8% of responses were from government departments, 8% from institutional archives and a further 6% from national archives. 6% of responses were from service or infrastructure providers and 5% from data centres. All suggested organisation types were represented by participants by the remaining 21% of responses (*see figure 1*).

16% of participants also specified an 'other' organisation type including manufacturers, public libraries, archives with a specialist remit, and organisations which are structured from two or more of the suggested organisation types.

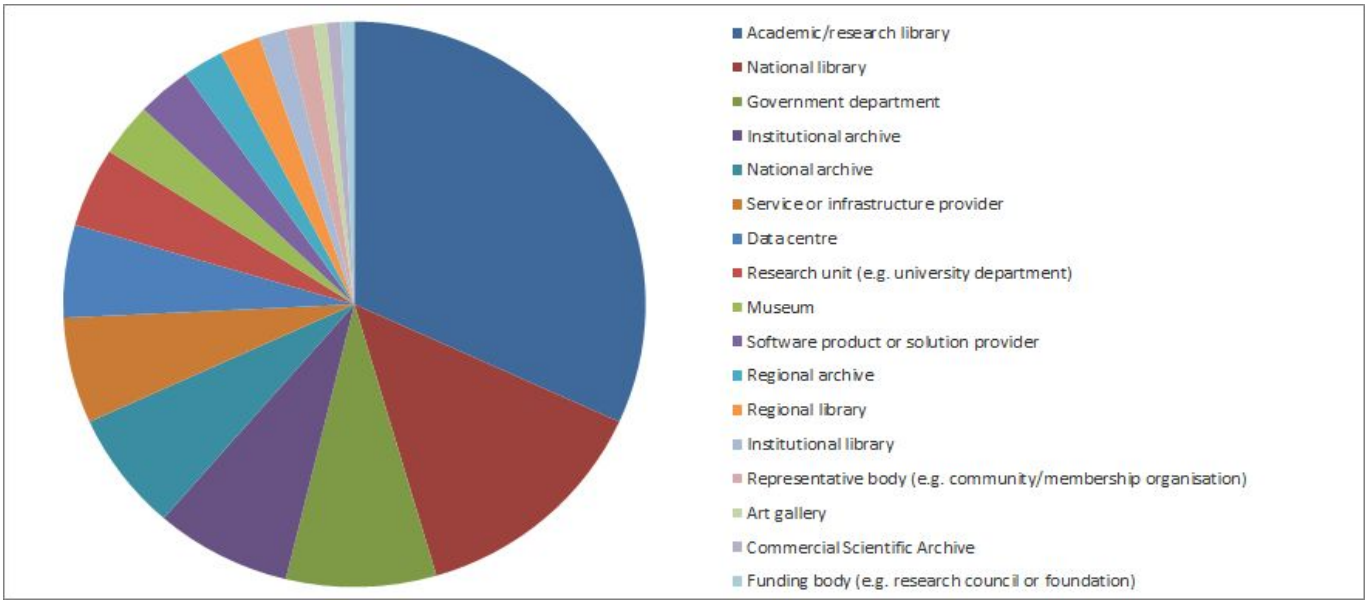

Figure 1: Organisation type

Academic/research library	32%
National library	14%
Government department	8%
Institutional archive	8%
National archive	7%
Service or infrastructure provider	6%
Data centre	5%
Research unit (e.g. university department)	5%
Museum	3%
Regional archive	3%
Software product or solution provider	2%
Regional library	2%
Art gallery	2%
Institutional library	2%
Representative body (e.g. community/membership organisation)	1%
Commercial Scientific Archive	1%
Funding body (e.g. research council or foundation)	1%

We received responses from 31 countries (*see figure 2*). The majority of responses, 39%, were from organisations based in the United States. 17% of responses came from organisations in the United Kingdom. Germany and Canada each account for 4% of responses, and Spain and Denmark each account for 3%.

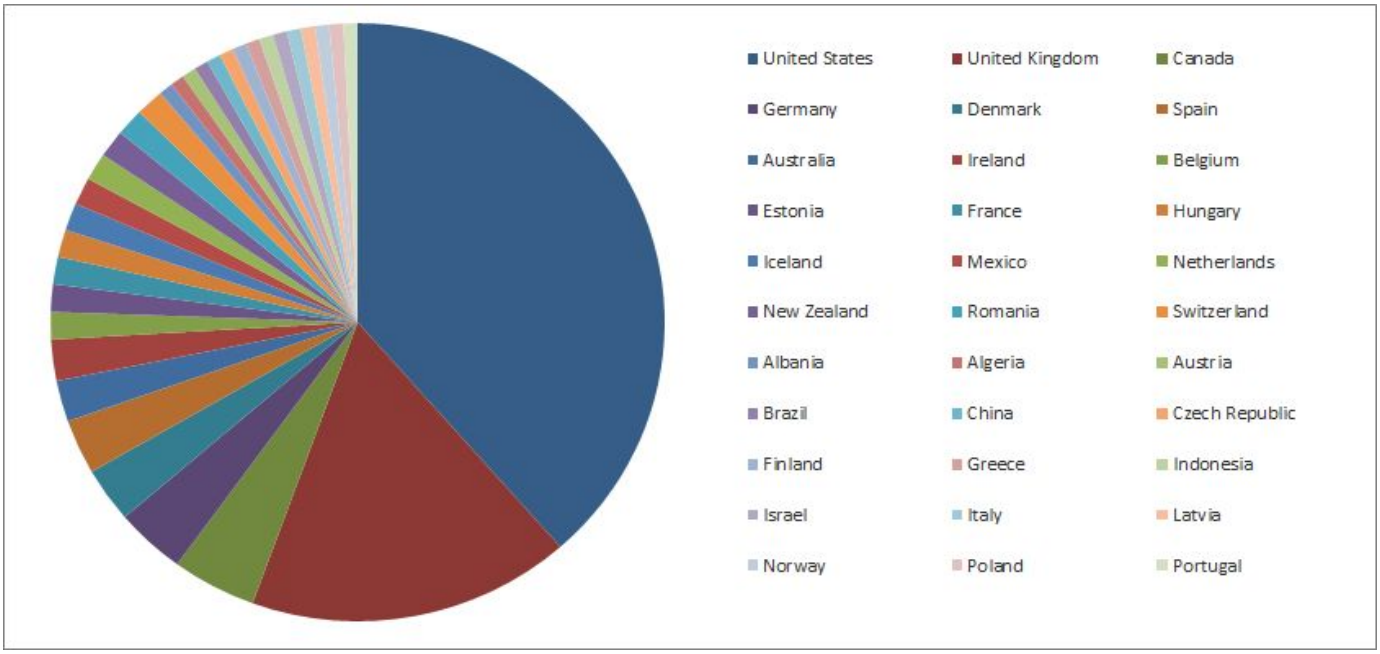

Figure 2: Geographical distribution

United States	39%	New Zealand	2%
United Kingdom	17%	Romania	2%
Canada	4%	Switzerland	2%
Germany	4%	Austria	1%
Denmark	3%	Brazil	1%
Spain	3%	China	1%
Australia	2%	Czech Republic	1%
Ireland	2%	Finland	1%
Belgium	2%	Greece	1%
Estonia	2%	Indonesia	1%
France	2%	Israel	1%
Hungary	2%	Italy	1%
Iceland	2%	Norway	1%
Latvia	2%	Poland	1%
Mexico	2%	Portugal	1%
Netherlands	2%		

3. Company information

3.1 Staffing

28% of the respondent organisations have 51-200 members of staff (s*ee figure 3)*. 23% have 200-1000 staff and 13% have over 1000 staff. 12% of organisations have 21-50 staff, 14% have 6-20 staff, and 10% have between 1-5.

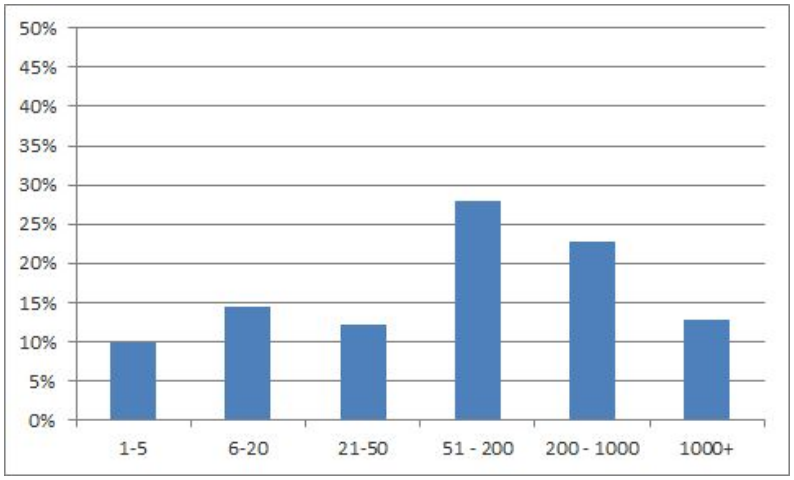

Figure 3: Number of staff

All of the suggested digital preservation roles are employed across participating organisations (*see figure 4*). Managerial or administrative roles are the most common amongst respondents with 72% employing staff in these roles. 66% of respondents have software developers or programmers in their digital preservation teams. 64% employ digital archivists or curators, and 59% employ system administrators.

We also asked about other digital preservation roles. Respondents commented that in some cases their organisation did not have a dedicated digital preservation team, or dedicated roles, but instead that digital preservation activities are carried out across different departments such as IT and research, or different people contribute to different stages of the digital preservation workflow.

Figure 5 reflects this showing the average full time employee (FTE) in each role across respondents. Software developers or programmers represent the highest role with an average of 1.51 FTEs.

Key finding
❏ The most common role for in digital preservation are managers and administrators ❏ Software developers or programmers represent the average highest FTE role in digital preservation ❏ The average number of staff for the majority of digital preservation roles is less than 0.8 FTE

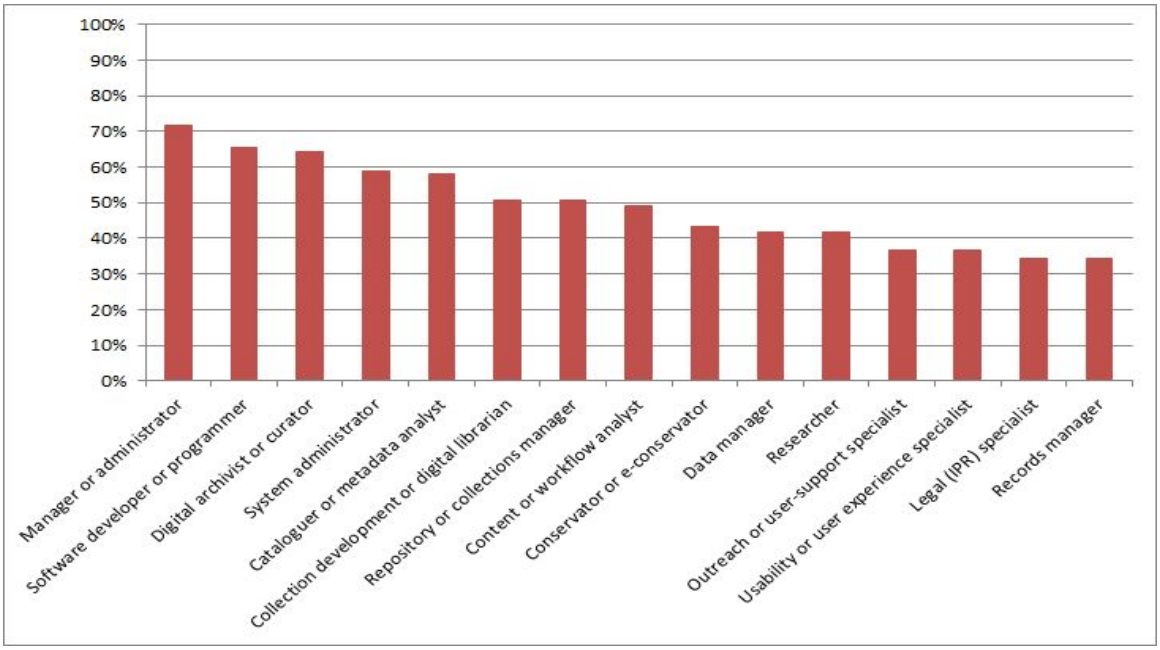

Figure 4: Proportion of respondents with each role in their team

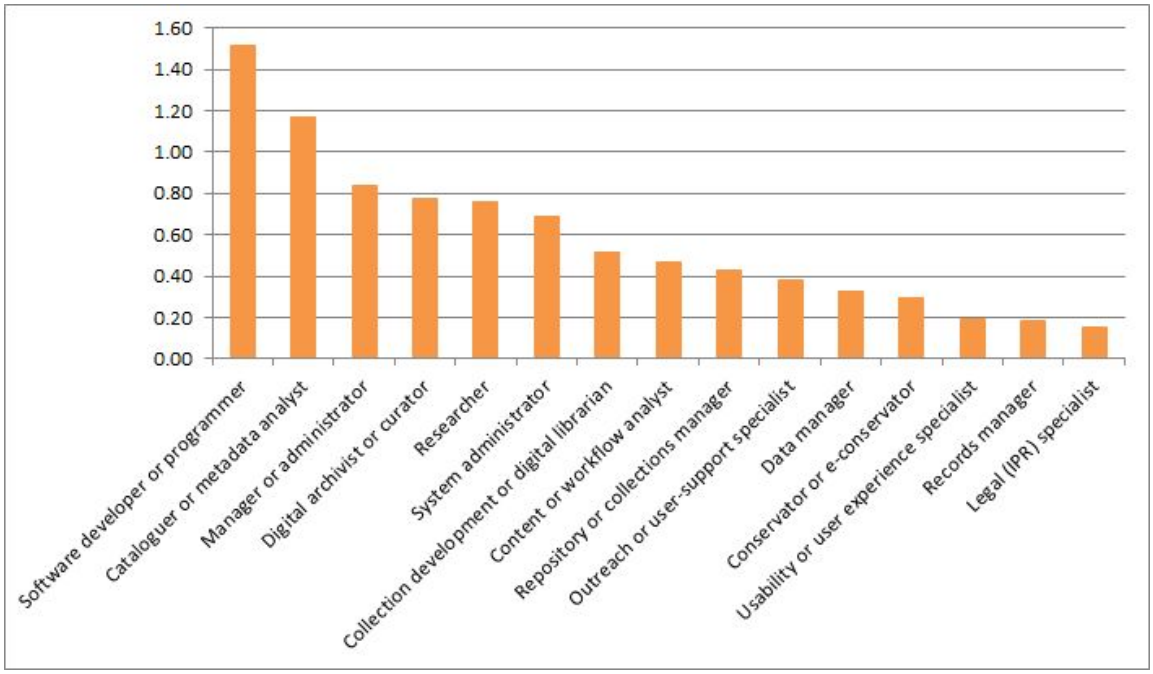

Figure 5: Average FTE for all reported roles

4. Activities and Approaches

All of the listed digital preservation activities are carried out by respondents in some way or another (*see figure 6*), although not all organisations carry out all activities.

The most widespread activities among respondents are digitisation, followed by bit preservation and software development or maintenance. Also common are metadata creation or extraction policy development , preservation planning and technology watch . Emulation appears to be less widely adopted as a preservation strategy than migration.

Carrying out digital preservation activities in-house is the most common approach across respondents. Digitisation, software development or maintenance and storage or bit preservation are the most common activities to be outsourced. Organisations make use of community solutions across all activities. Technology watch, software development or maintenance and preservation research are the most common activities to rely on community solutions.

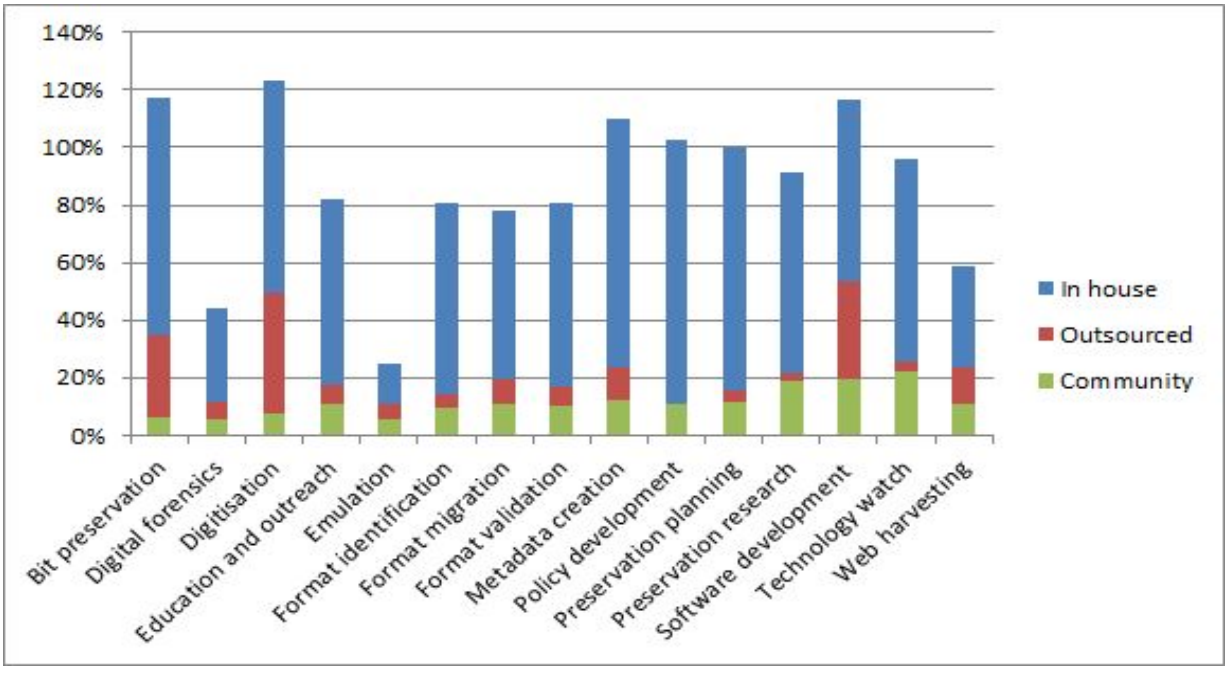

Figure 6: Delivery of digital preservation activities

Key findings
❏ Most of the organisations that responded, carry out their core digital preservation activities in house. ❏ Digitisation, software development and bit preservation are the most heavily outsourced activities. ❏ Community solutions make a contribution across all activities.

4.1 Collections

Content profile

Of the collection-holding respondents, 90% hold images (*see figure 7*). 88% hold unstructured documents, 74% collect audio formats, 74% have structured documents and 70% hold video formats. 58% collect container formats, and 49% hold databases/database records. Content types are held by 20% or less of collection-holding respondents include 3D formats, digital artworks and hardware or environments.

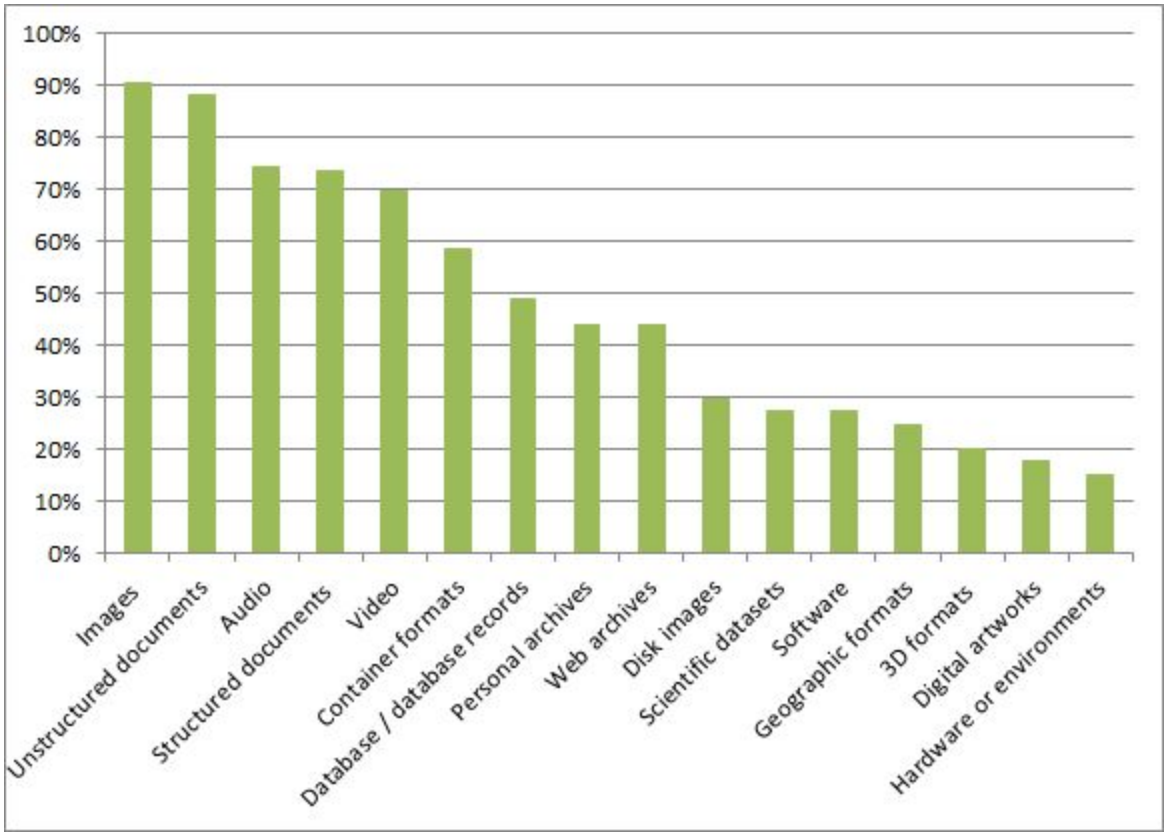

Figure 7: Content types held in collections

It appears that organisations are still prioritising the management of formats with clear analog counterparts where the challenges are better understood. Despite recognising the importance of more digitally native formats, organisations have not collected these at the same scale yet. For example, respondents to the 2014 survey report lower usage rates of tools for databases and websites (e.g. SIARD and Heritrix) than of tools for more traditional formats.

Key findings
❏ Images, documents, audio and video formats are most commonly held by respondents, showing very similar results to the 2009 Planet survey respondents. ❏ The representation of audio and video formats grew the most between the 2009 and 2014 surveys, from approximately 50% to 75% in both cases. ❏ In contrast, the growth in the use of website and database formats expected in 2009 failed to materialise by 2014. By 2019, respondents expect the highest increase in representation for databases and websites.

Storage Growth

11% of respondents have over 1 petabyte of storage capacity (*see figure 8*). 16% have between 101 terabytes and one petabyte and 28% have between 11 and 100 terabytes of storage. 30% of respondents have between 1 and 10 terabytes, and 15% have under a terabyte.

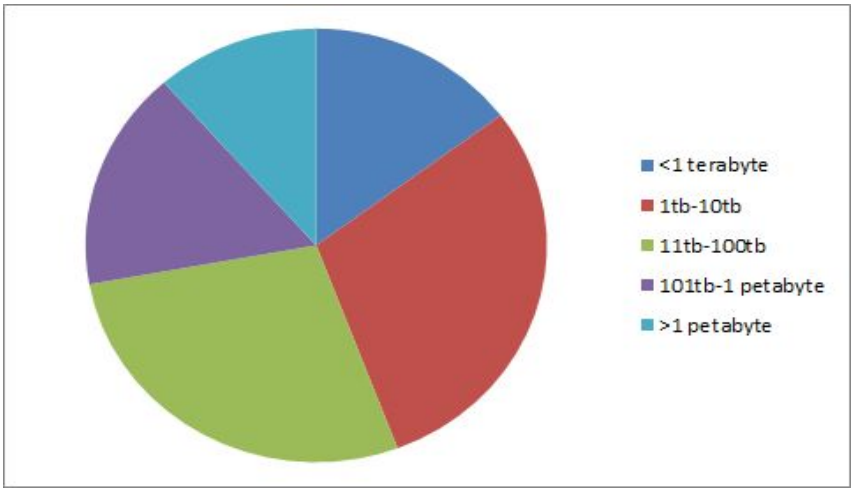

Figure 8: Current storage capacity

23% of organisations expect their storage capacity to grow between 1-10% over the next 12 months (*see figure 9*) and 29% expect it to grow between 11-25%. 15% expect growth between 26-50%, four expect growth between 51-75% and a further 15% expect 75-100% growth. 10% of respondents do not know

whether their storage capacity will increase over the next 12 months, and 5% do not expect it to increase. No organisations thought their storage capacity would decrease.

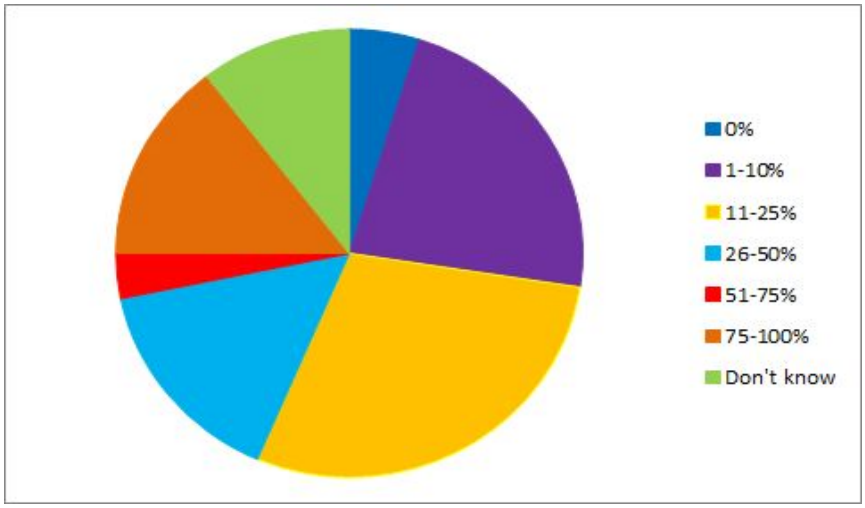

Figure 9: Expected growth

The 2009 Planets survey asked organisations how much data they stored at the time and to predict how much they would need in two, five and ten years time.

Holdings have not increased as quickly as expected.

- 56% of organisations surveyed by Planets anticipated holding over 100TB by 2014. The OPF survey in December 2014 shows only 27% of organisations currently hold this volume of data.
- 57% of organisations surveyed by OPF hold from 1-100TB of data; in 2009 the Planets survey showed 58% of organisations with this volume of data, an almost identical finding.
- Only 36% of the Planets survey respondents anticipated holding 1-100TB by 2014.

It is unclear why holdings have not increased as expected. Potential reasons include:

- Digital collections growing slower than expected
- Repository systems being unable to scale with demand
- Data producers using other storage options
- Adequate preservation solutions not existing for formats such as databases
- Lack of funding for the heritage and digital preservation sectors

5. Infrastructure

5.1 Production environments

The majority of infrastructure is hosted locally by respondents (*see figure 10*). Some organisations additionally outsource their infrastructure, host it in the cloud or run applications as part of a consortium. Hosting infrastructure in the cloud is the least common method across each of the applications.

Content is the most common type of infrastructure to be outsourced, with 20% if organisations storing their content with external providers. 14% outsource their repository systems, 12% outsource their metadata systems and 4% outsource their processing systems.

19% of organisations host their metadata systems as part of a consortium. 18% host their content, 15% host their repository system and 4% host their processing systems with a consortium.

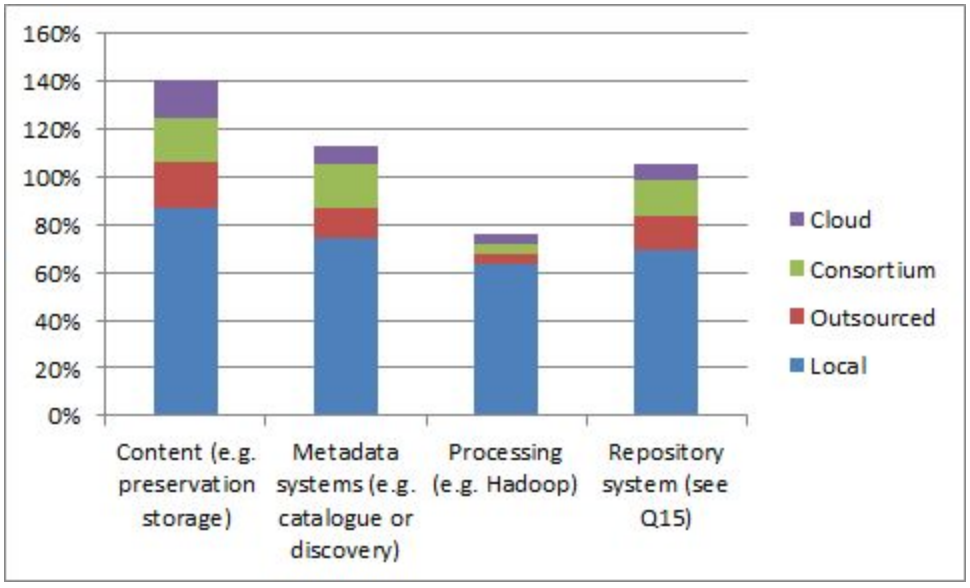

Figure 10: Infrastructure

5.2 Operating systems

Linux is the most widely-used server, in use by 64% of respondents, followed by Windows with 53% respondents using it in their organisations (*see figure 11*). 26% of organisations use Unix, and 5% use OSX. On the desktop, Windows predominates in use by 83% respondents. 36% respondents use OSX, 34% use Linux and 12% use Unix.

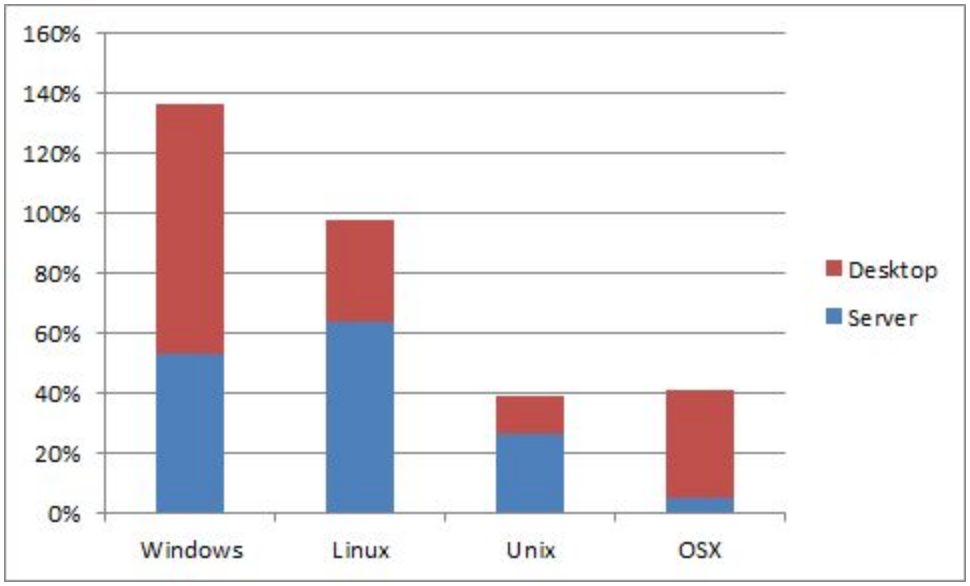

Figure 11: Operating systems

6. Use of open source

78% of respondents use a mixture of open-source and proprietary technology. 10% use entirely open-source technology and only 11% do not use any open-source technology (*See figure 12*).

19% maintain or lead collaborative open-source technology projects, and 33% contribute to collaborative open-source technology projects. Just under half of respondents (48%) do not contribute to collaborative open-source technology projects. (*See figure 13*)

In comparison to the Planet survey from 2009, 13% used all open source technology, 14% used proprietary solutions, 57% used a mix of the two and 16% had not decided what to use. When asked what they thought their use of open source vs proprietary technology would be in future, 25% had not decided. Expectations of use of proprietary solutions decreased significantly to only 2%, and use of open-source software remained similar at 14%.

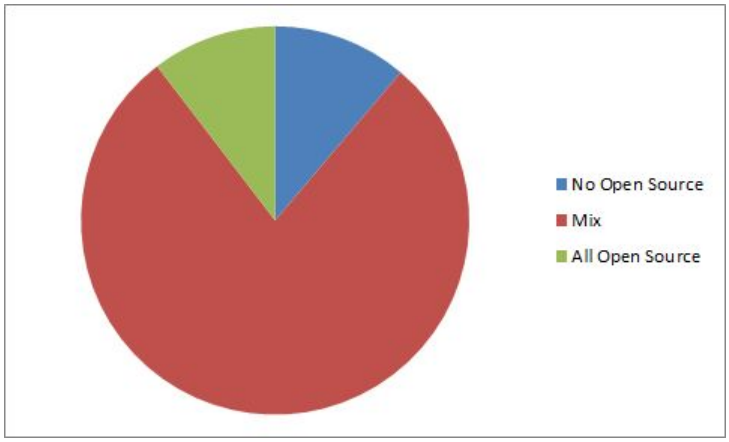

Figure 12: Use of open source technology

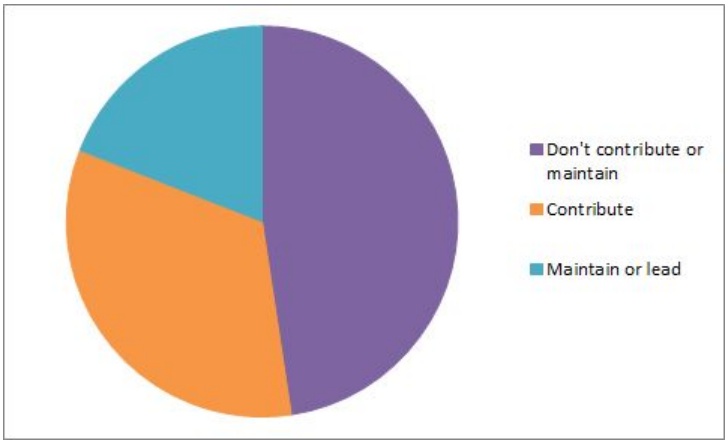

Figure 13: Contribution to open source projects

Key findings
❑ Open source software is used by 88% of respondents, showing its importance in the digital preservation community. ❑ Only 12% of organisations make no use of open source software at all.

7. Repository systems

Most respondents have a bespoke (in-house) repository system in production (*see figure 14*). DSpace and LOCKSS are each used in production by 19% respondents and Fedora used by 18% respondents. Fedora with Hydra is the most common system under evaluation by respondents, with 11% of organisations currently assessing it for their needs. Fedora alone is under evaluation by 8% of organisations and Fedora with Islandora by 11%.

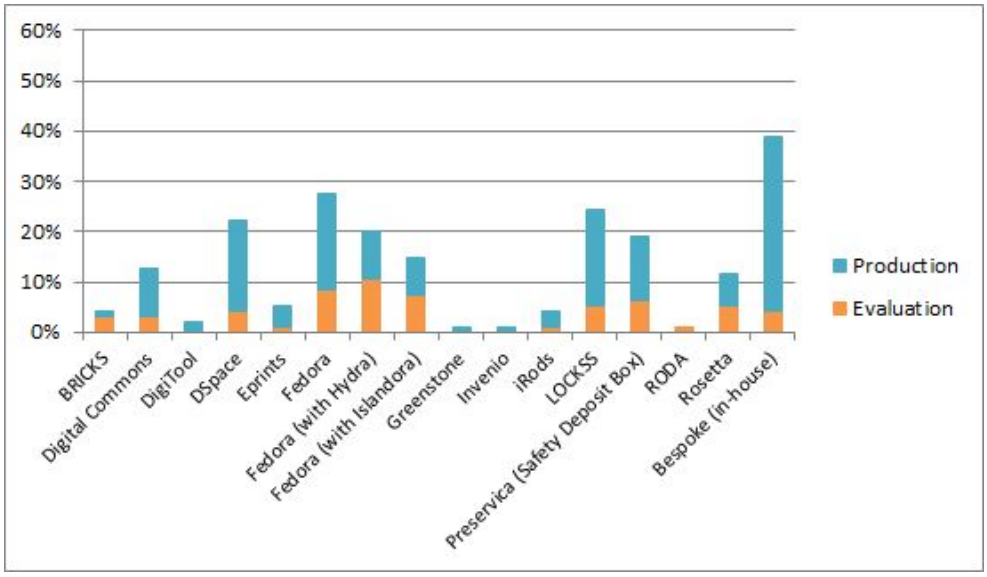

Figure 14: Use of repository systems

We also asked about the importance of these systems (*see figure 16*), where '1' indicates low importance, and '3' high importance[1]. A bespoke (in-house) system is rated as '3' by the majority of organisations. Fedora with Hydra, Fedora and LOCKSS are also rated with high importance.

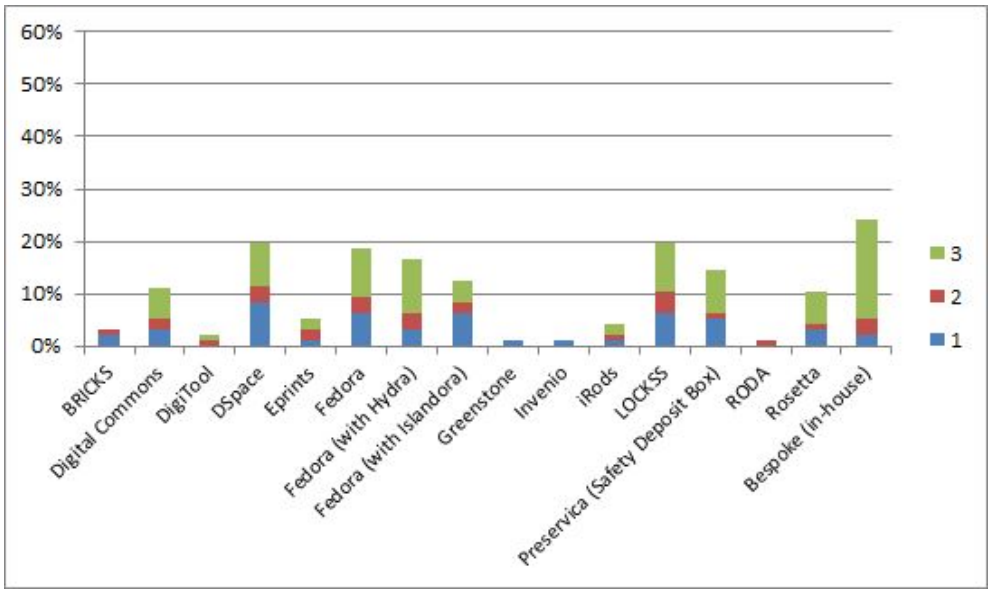

Figure 16: Importance of repository systems

[1] The results rated 1-5 have been grouped 1-2 (1 – low importance), 3 (average importance) and 4-5 (3 – high importance).

8. Workflow tools

45% of respondents use bespoke (in-house) workflow tools in production (*see figure 17*). 31% have workflow tools packaged with their repository system in production. Archivematica is the most popular workflow tool under evaluation (22%).

Respondents were also asked to rate the importance of these workflow tools (*see figure 18*). Workflow tools packaged with a repository system are rated as '3' important by the majority of respondents (27% of organisations), followed by bespoke (in-house) systems 25%). 13% of organisations rated Archivematica as '3' important.

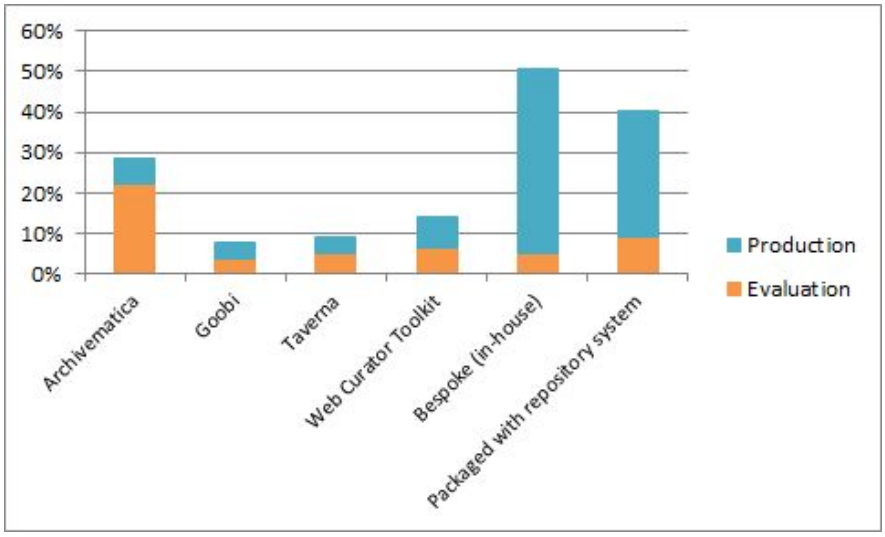

Figure 17: Use of workflow tools

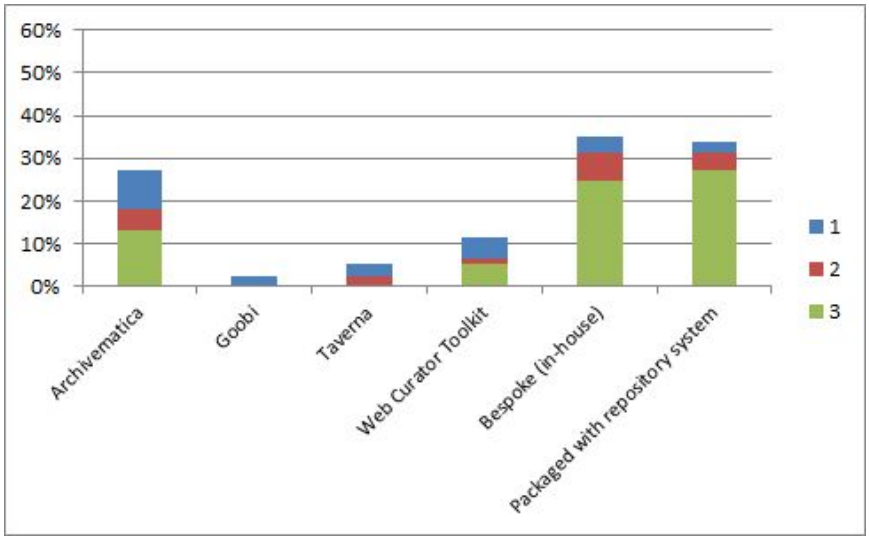

Figure 18: Importance of workflow tools

9. Technology

9.1 Tools

Two types of tool were included in the survey: dedicated digital preservation tools (such as DROID and JHOVE) and other tools which are used for digital preservation (such as Tika and PDFBox). All tools listed are currently either in production or are under evaluation by at least one respondent.

ImageMagick is the most commonly used in production (28% of organisations) and under evaluation by a further 5% (*see figure 19-20*) It is the third highest ranked in terms of importance (*see figure 21-22*). JHOVE and DROID are in production by 23% of respondents and under evaluation by 7% and 14% respectively. JHOVE is rated as most important tool equally with DROID.

The tool which is most commonly under evaluation is BitCurator by 18% of organisations. It is rated as '3' high importance by 13% organisations.

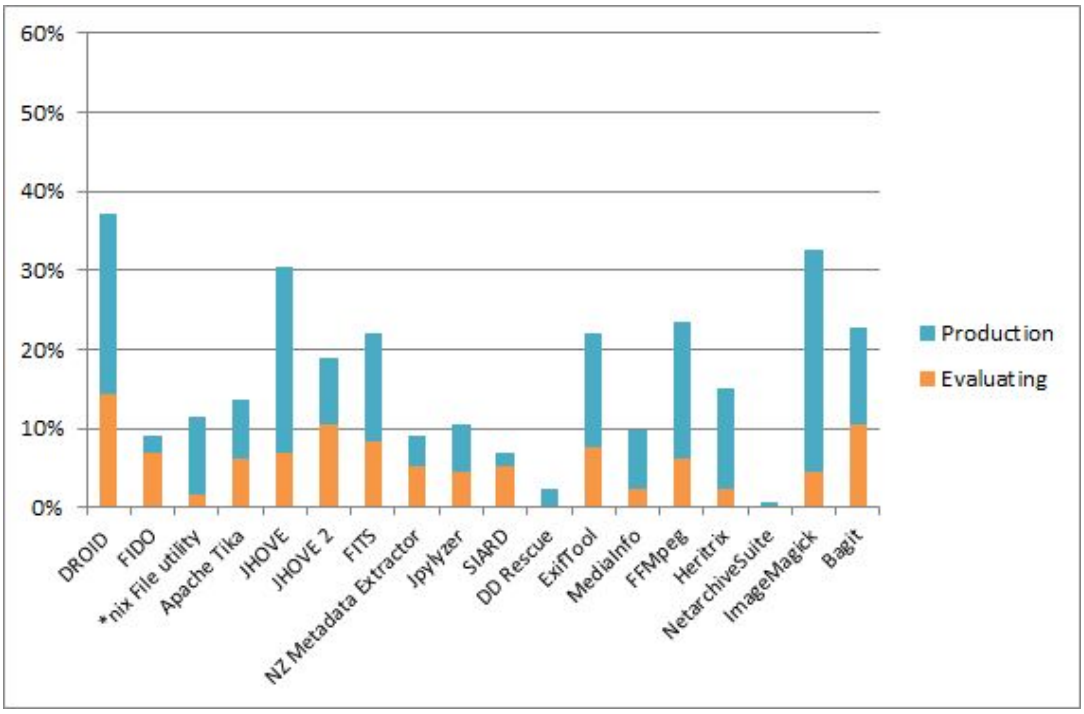

Figure 19: Use of tools

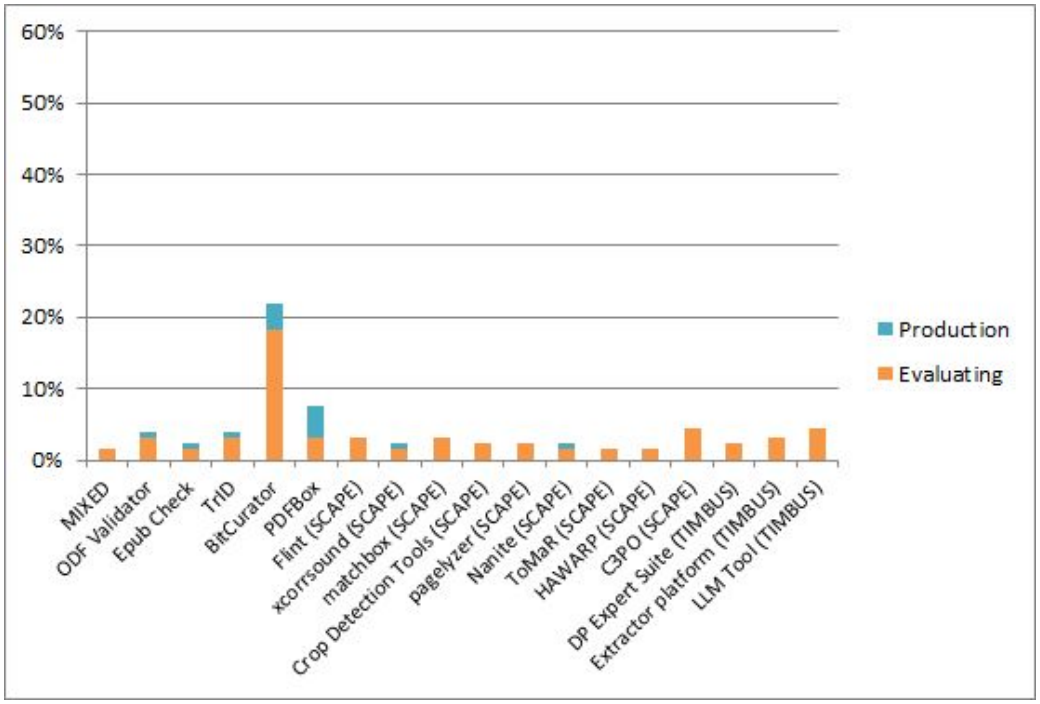

Figure 20: Use of tools continued

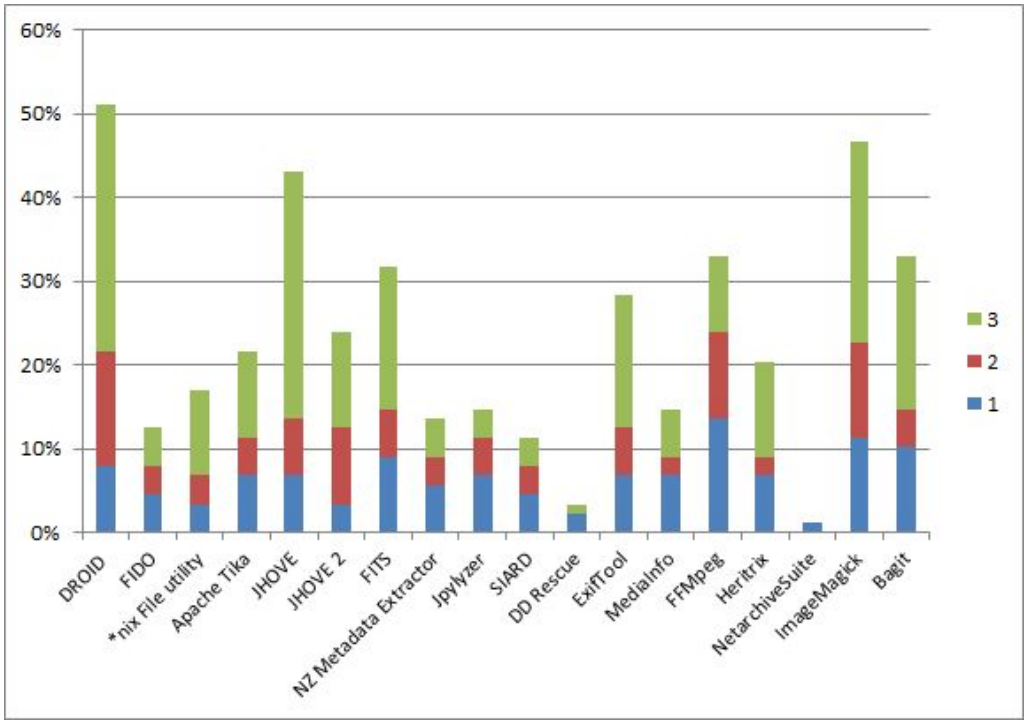

Figure 21: Importance of tools (1-low, 3-high)

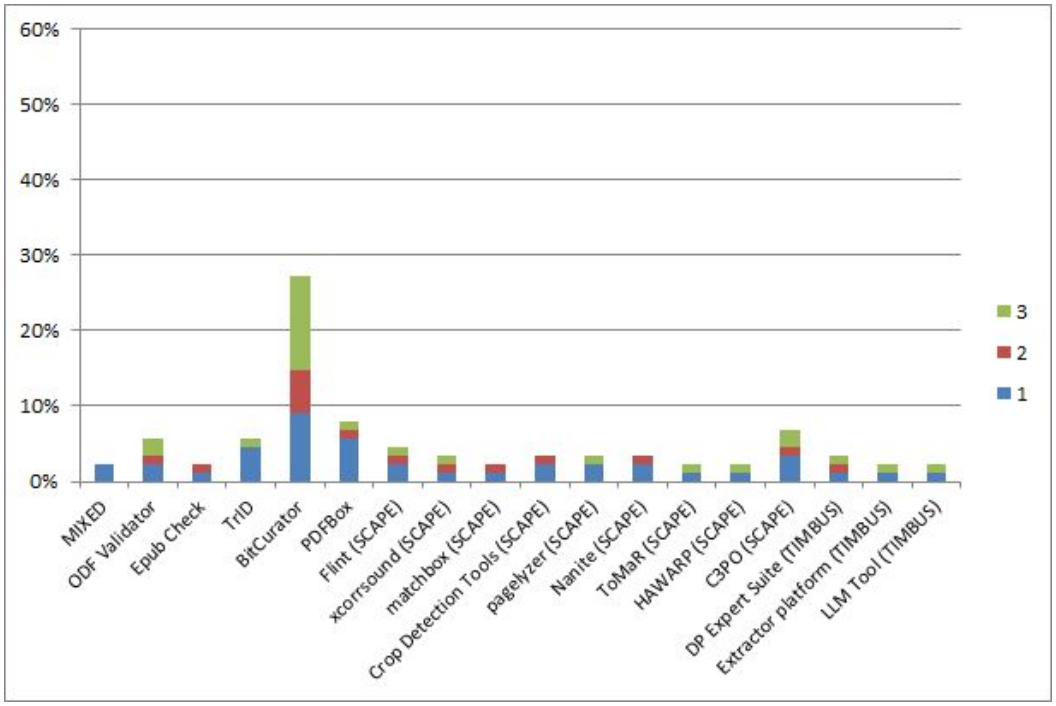

Figure 22: Importance of tools (1-low, 3-high) continued

9.2 Services

Of all the services, PRONOM is most commonly used in production (23% of organisations), is under evaluation by the highest number of respondents (13% of organisations) and of highest importance (20% of organisations).

The OPF Knowledge Base (wiki) is next most commonly used with 5% of organisations using it in production and a further 5% evaluating it. We will continue to use the wiki as a community space, migrating its hosting to new infrastructure and curating its content to make existing materials more visible, for example through interest groups.

COPTR is being used by 3% of organisations and is being evaluated by a further 9%. OPF will continue to host COPTR, and the digital preservation Q&A site, which is used by 2% and under evaluation by another 4%.

bwFLA (emulation as a service) is used in production by 2% and is under evaluation by a further 5% OPF will continue to work with the University of Freiburg, as an affiliate member, to further the maturity of the service and support its evaluation and adoption.

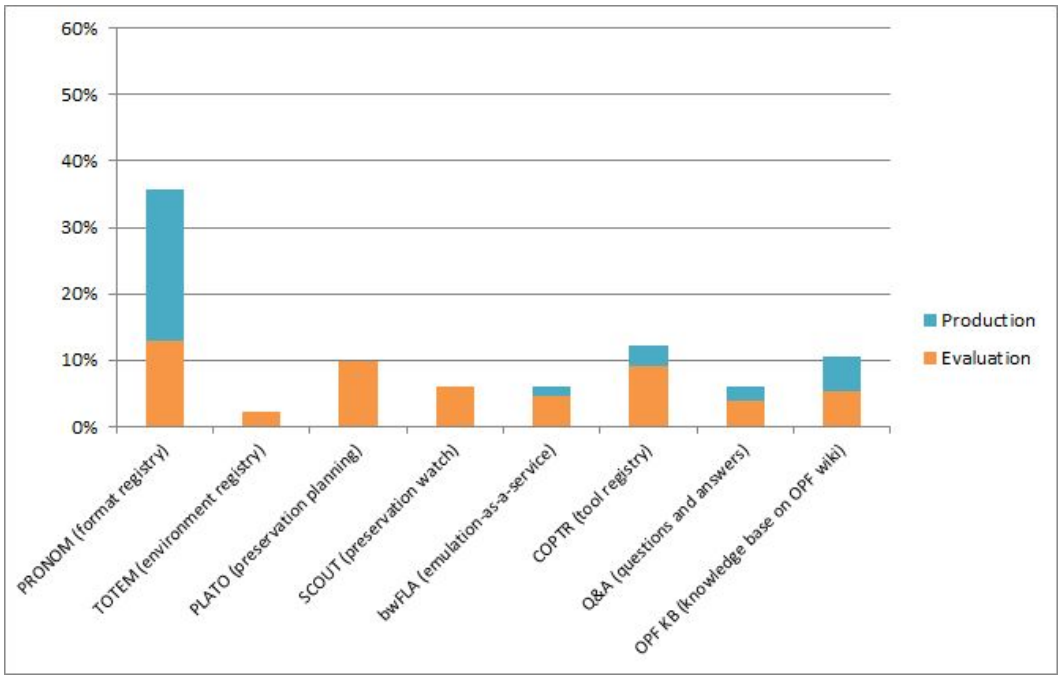

Figure 23: Use of services

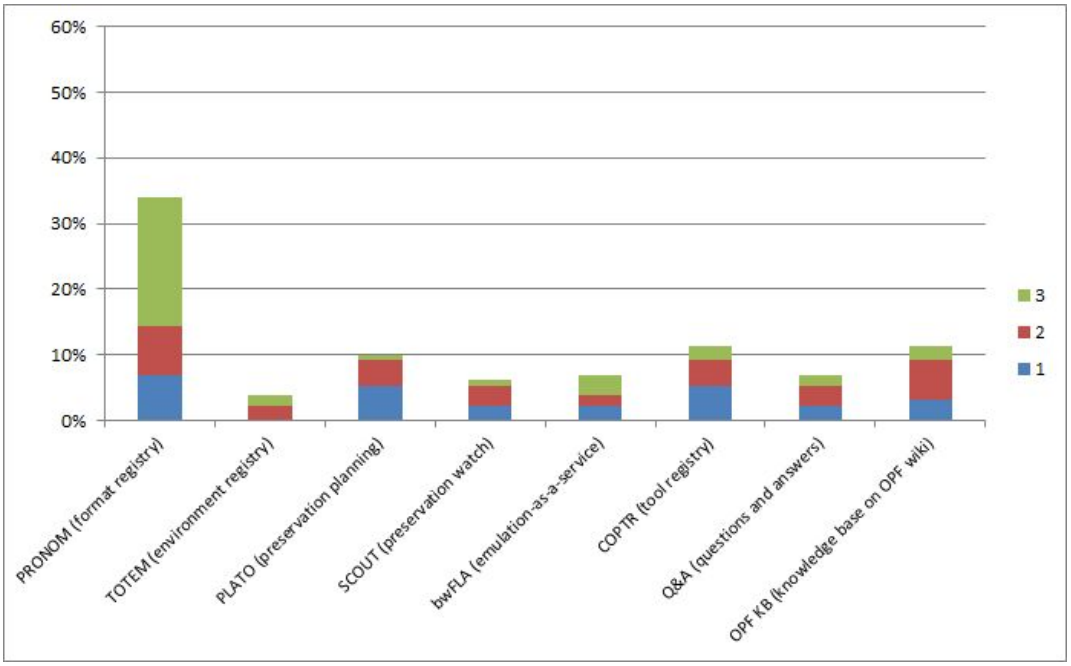

Figure 24: Importance of services (1-low, 3-high)