2019-20 DIGITAL PRESERVATION COMMUNITY SURVEY

Open Preservation Foundation

Findings report

CONTENTS

Introduction

The Open Preservation Foundation leads a collaborative effort to create, maintain and develop the reference set of sustainable, open source digital preservation tools and supporting resources.

Digital preservation is a growing community and the more we know about it, the more we can do to collectively support its development. Our Digital Preservation Community Survey was designed to capture community approaches to digital preservation and enable us to build a knowledge-base about the tools and resources relied on by memory institutions carrying out digital preservation work today. The survey ran between November 2019 and February 2020, and the results provide a detailed picture of the digital preservation landscape today.

This report is based on a simple interpretation and understanding of the results. Where possible, we have drawn comparisons to the results of the 2015 community survey and a 2009 survey conducted as part of the PLANETS project. The anonymised data has been made openly available and we encourage you to carry out your own analysis and benchmark your practice with others.

Most respondents answered all of the questions, but figures for questions with less than a 100% response rate include the number of responses for reference (e.g. 'n=87' would indicate that the question received 87 responses).

Guidance notes, appendices, reports and raw data from our previous surveys can be found at www.openpreservation.org/surveys.

Respondents

We received responses from 98 organisations in 31 countries. Of these countries, 51% are in Europe, 35% in North America and 5% in Africa. The remaining 9% of responses were split between countries in South America, Asia and Australasia.

Of the 132 responses received to our 2015 community, we received a similar percentage (47.8%) from European countries, 41.2% from North America, 3.7% from Australasia and 2.2% from South America. Responses from African countries are up from 1 in 2025 to 5 in 2020.
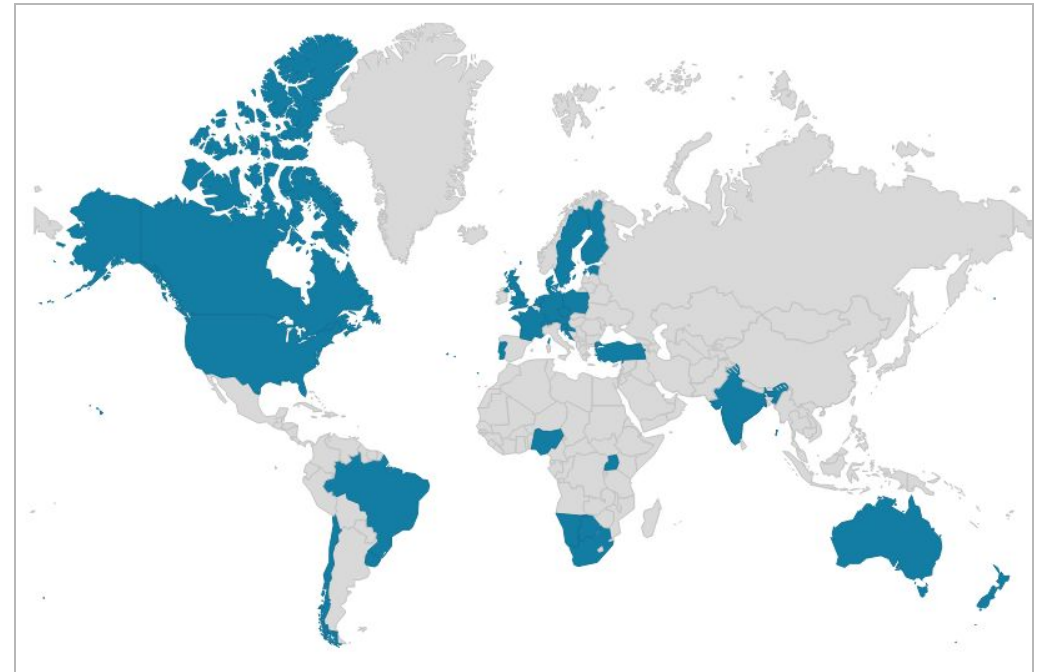

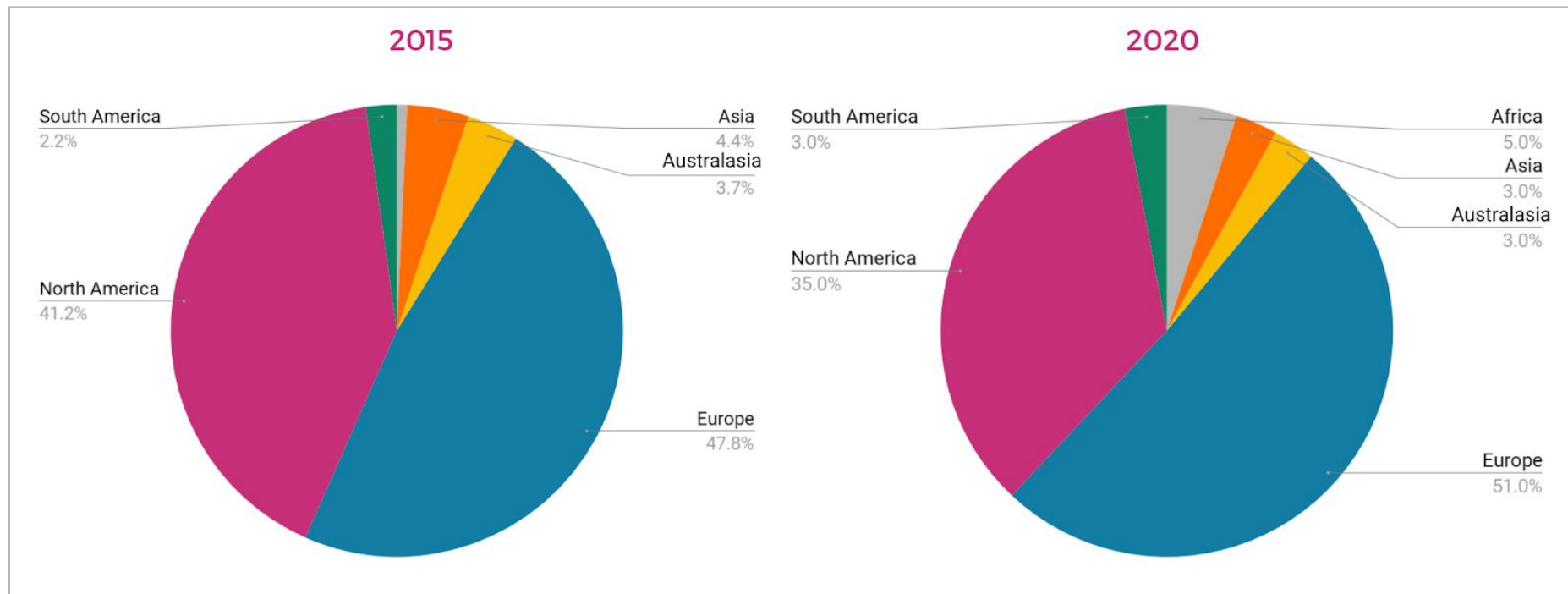*Fig. 1. Map showing respondents' countries*

Fig 2. Comparison of responses by continent in 2015 and 2020

The majority of responses came from Academic/research libraries (30.6%), national libraries (12.7%), national archives (7.5%) and museums (6.7%). 3% of respondents did not feel that any of the options on the list accurately reflected their type of organisations. The 'other' types of organisation included 'International Organisation Archive', 'Photographic archive', 'Archive within a public broadcasting organization' and 'Discipline specific data archive'. Organisation types with less than 1% representation within the results (and therefore not labelled on the chart) are representative body, municipality archive, funding body, and commercial scientific archive.
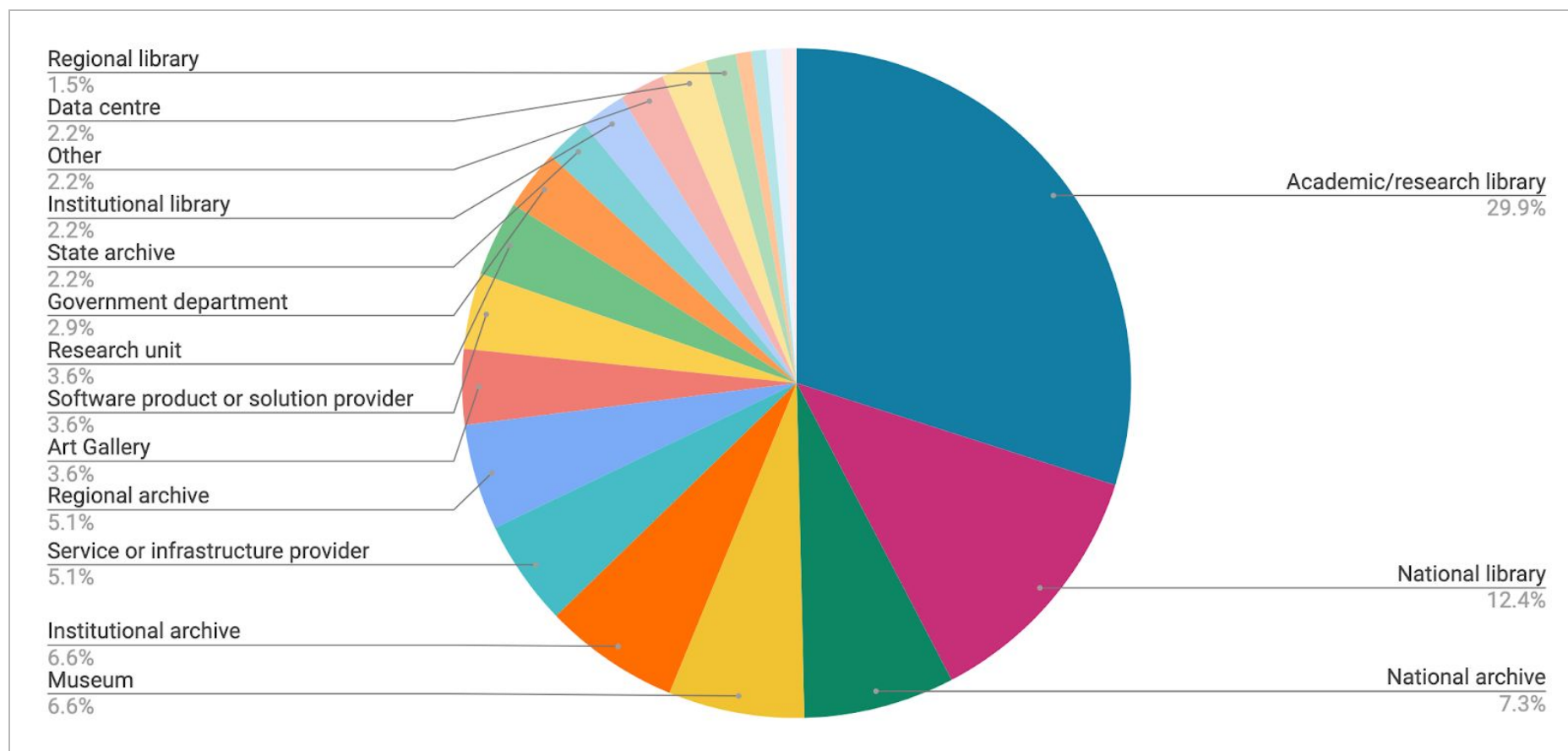

Fig. 3. Responses by type of organisation

These results are similar to those of the 2015 survey, where 32% of responses came from academic and research libraries, 14% came from national libraries and 7% came from national archives. However, the percentage of responses we received from museums more than doubled from 3% in 2015 to 6.7% in 2020.

Staffing

Respondents were asked how many members of staff are employed by their organisation, and the results demonstrate varied organisational capacity (fig. 4).

We classified small organisations as those with between zero and fifty members of staff. 16% of respondents' organisations have between just 6-20, 11% have between 21-50 staff, and 3% have fewer than 5 employees.

Mid-sized organisations were the most common, making up 43% of responses overall. 23% of organisations have between 201-500 employees and 20% have between 51 and 200.

Large organisations (501-1001+) made up 24% of responses: 16% have over 1001 members of staff and 9% have 501-1000.
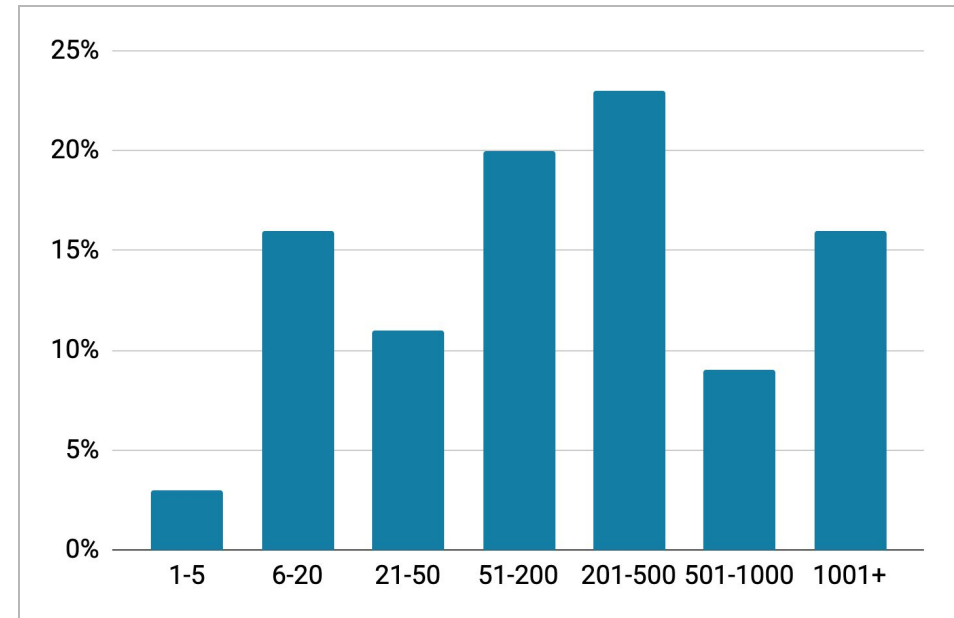

Fig. 4. Number of staff

We asked respondents to indicate, to the best of their knowledge, which job roles play a direct part in their organisation's digital preservation activities (fig. 5), and how much time is spent on these activities in Full Time Equivalent (FTE) (fig. 6).

A list of suggested roles was provided. These roles are not necessarily limited to dedicated digital preservation staff, teams or units, but refer to all employees who contribute to digital preservation activities across the organisation.

As in 2015, each of the suggested roles is employed by at least one organisation and the most common roles are director, manager or administrator and cataloguer or metadata analyst, each of which is employed at 68% of organisations.
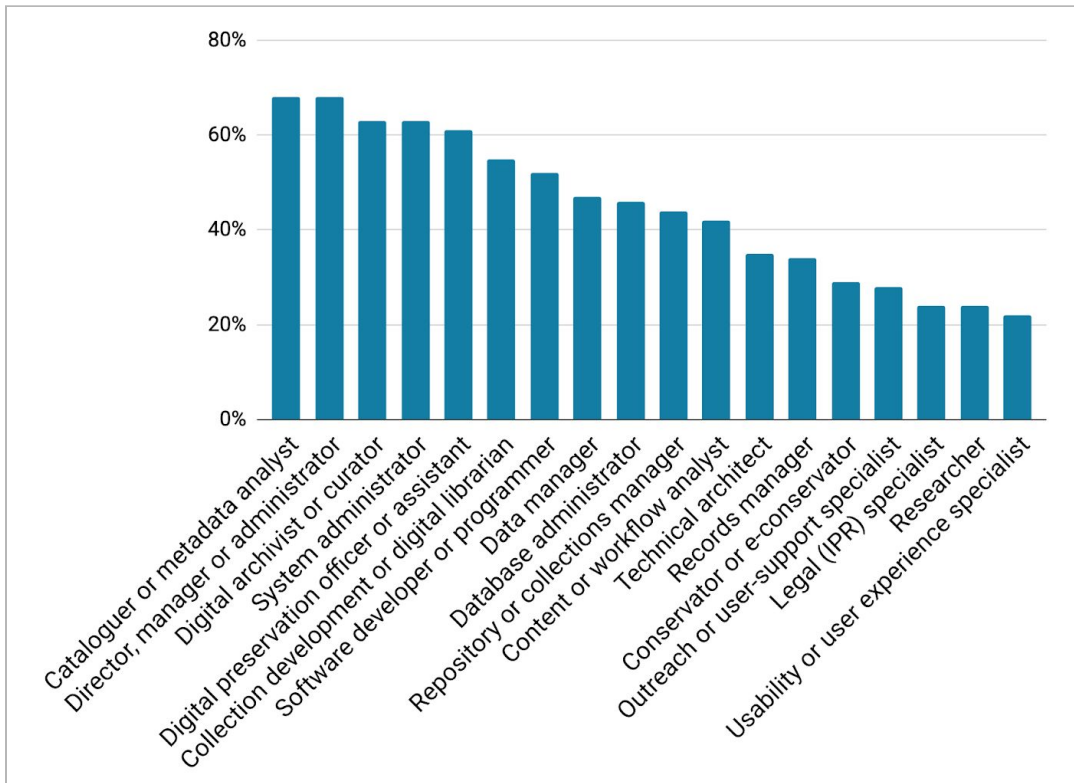

Fig. 5. Percentage of organisations that employ each specified role

Popular roles employed at over 50% of organisations are:

- Software developer or programmer
- Collection development or digital librarian
- Digital preservation officer or assistant
- System administrator
- Digital archivist or curator

Participants were invited to specify any additional roles not covered by the list provided - these included metadata strategist, IT manager and digitizer.

Less than a quarter of organisations employ usability or user experience specialists, researchers or legal specialists that contribute to digital preservation. However, although only 24 organisations hire researchers, this role has the highest average FTE and is employed predominantly across academic and research libraries and research units.
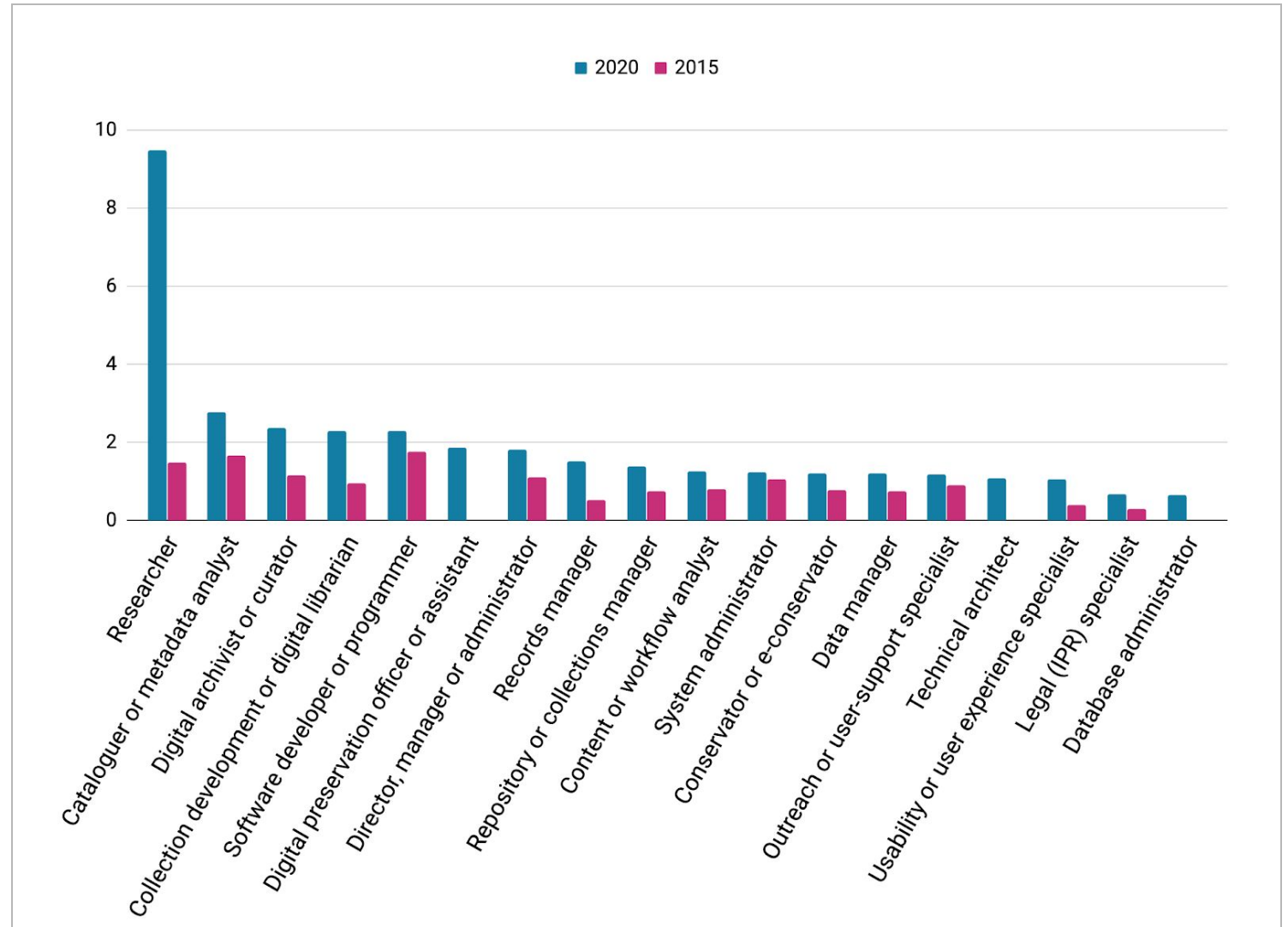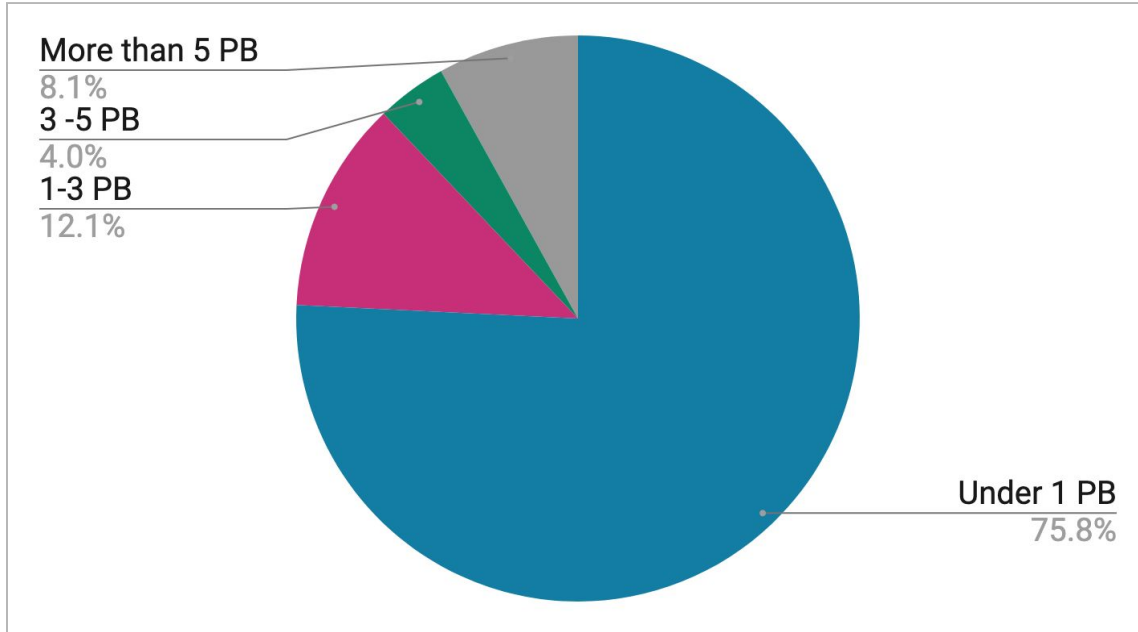

Fig. 6. Comparison of average FTE per role in 2015 and 2020

The average FTE across digital preservation roles in 2015 was 0.8 FTE. Our data indicates that this has now more than doubled to 1.96 FTE, suggesting growth in organisational capacity for digital preservation activities over the five year period.

Collections

More than 5 PB
8.1%

3 -5 PB
4.0%

1-3 PB
12.1%

Under 1 PB
75.8%

We asked respondents about the amount of data currently held by their organisation (see fig. 6). Approximately three quarters (75%) of organisations have under one petabyte (PB) of storage. The remaining 25% have more that one petabyte, with 12 organisations reporting between 1-3 PB of storage, 4 organisations between 3-5 PB. 8% of respondents including national libraries and archives, academic libraries and software or product solution providers have more than 5PB of storage capacity.

76 respondents gave an estimated total number of digital objects in their collection. Their responses range from just 212 to 1 billion.

Fig. 7. Current amount of data held by percentage of respondents (n=97)

When asked how much they expected their volume of content to grow in the next year, 10% of respondents were uncertain. No respondents expect their digital collections to decrease in size or stay the same in the next twelve months. 90% anticipate an increase, with 33% predicting an increase of between 1-10% and 33% predicting an increase of between 11-25%. 10% of organisations expect a growth of between 26-50%, with just 2% of respondents anticipating growth of between 51-75%.
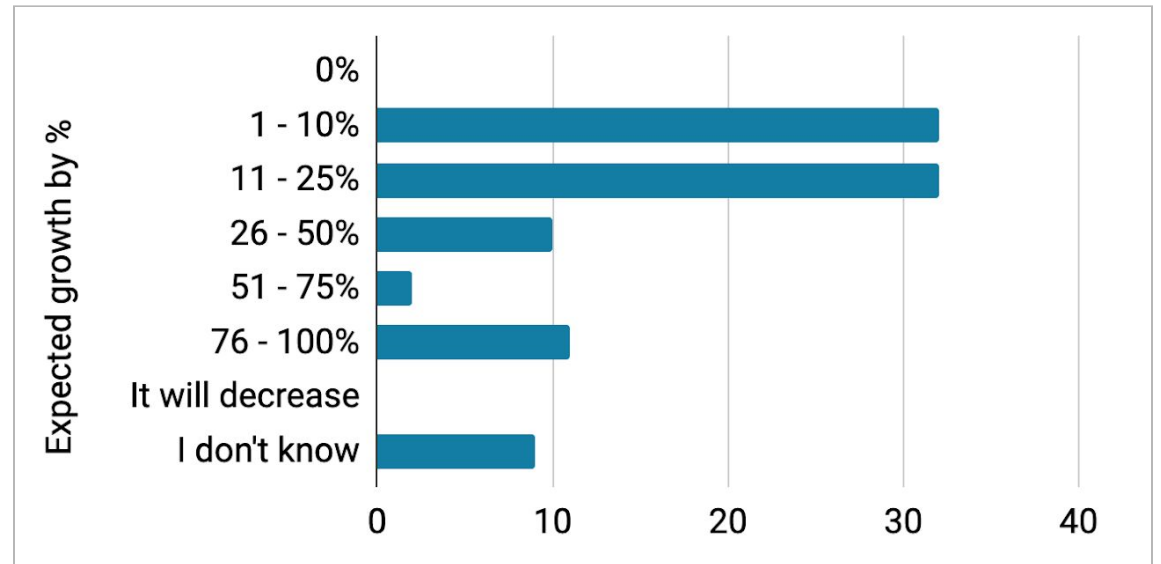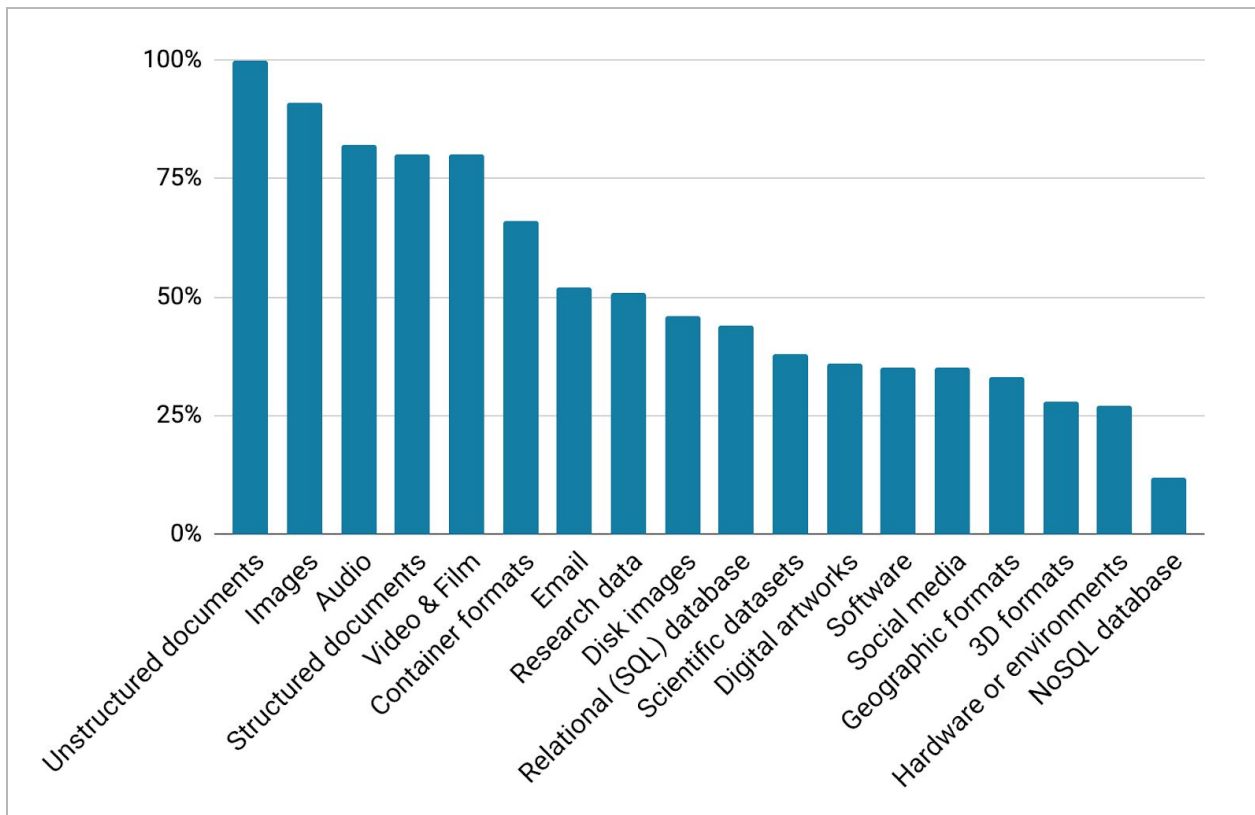

Fig. 8. Expected growth within 12 months

Fig. 9. Types of content by percentage

We asked respondents for a simplified analysis of their core collection areas to enable us to gauge the diversity of the content held in their preservation care.

All of the respondents' collections (100%) contain unstructured documents such as ebooks, ejournals, office documents, and PDFs. This demonstrates an increase of 13.6% in popularity in the past five years.

Another very common content type is images (2D, still images), which are held by 91% of organisations up from 90% since 2015). During the time between our PLANETS survey in 2009 and the community survey in 2015, the representation of audio and video formats in respondents' collections grew from 50% to 75%. Although figures for each of these formats have now risen again to 82% and 80% respectively, growth has slowed somewhat in the past five years.

Structured documents (e.g. spreadsheets, CSVs, XML) have also grown in representation from 74% to 80%. Neither social media nor email was included in the list of suggested content-types in 2015, which may reflect a surge in collecting these kinds of content.

2015 survey respondents expected an increase in the representation of databases across collections by 2019. This has failed to materialise, with database representation in fact falling from 48% in 2015 to 44% today. Less than half of all organisations have the following content types in their collections:

- Disk images (46)%
- Relational (SQL) Database / database records (44%)
- Scientific datasets (38%)

- Software (35%)
- Social media (35%)
- Geographic formats (33%)
- 3D formats (28%)

- Hardware or environments (27%)
- Digital artworks (26%)
- NoSQL database (12%)

Digital Preservation Activities

We asked about the activities being carried out as part of respondents' digital preservation programmes, including whether their organisation is researching, developing the capacity to implement, or actively carrying out an activity in a production workflow.

A list of suggested activities was provided, all of which are carried out in some organisations in one way or another.

The most common activities being carried out in production by over half of the organisations are digitisation, metadata creation/extraction, format validation, storage or bit preservation and format identification. This aligns with the results of the 2015 community survey.
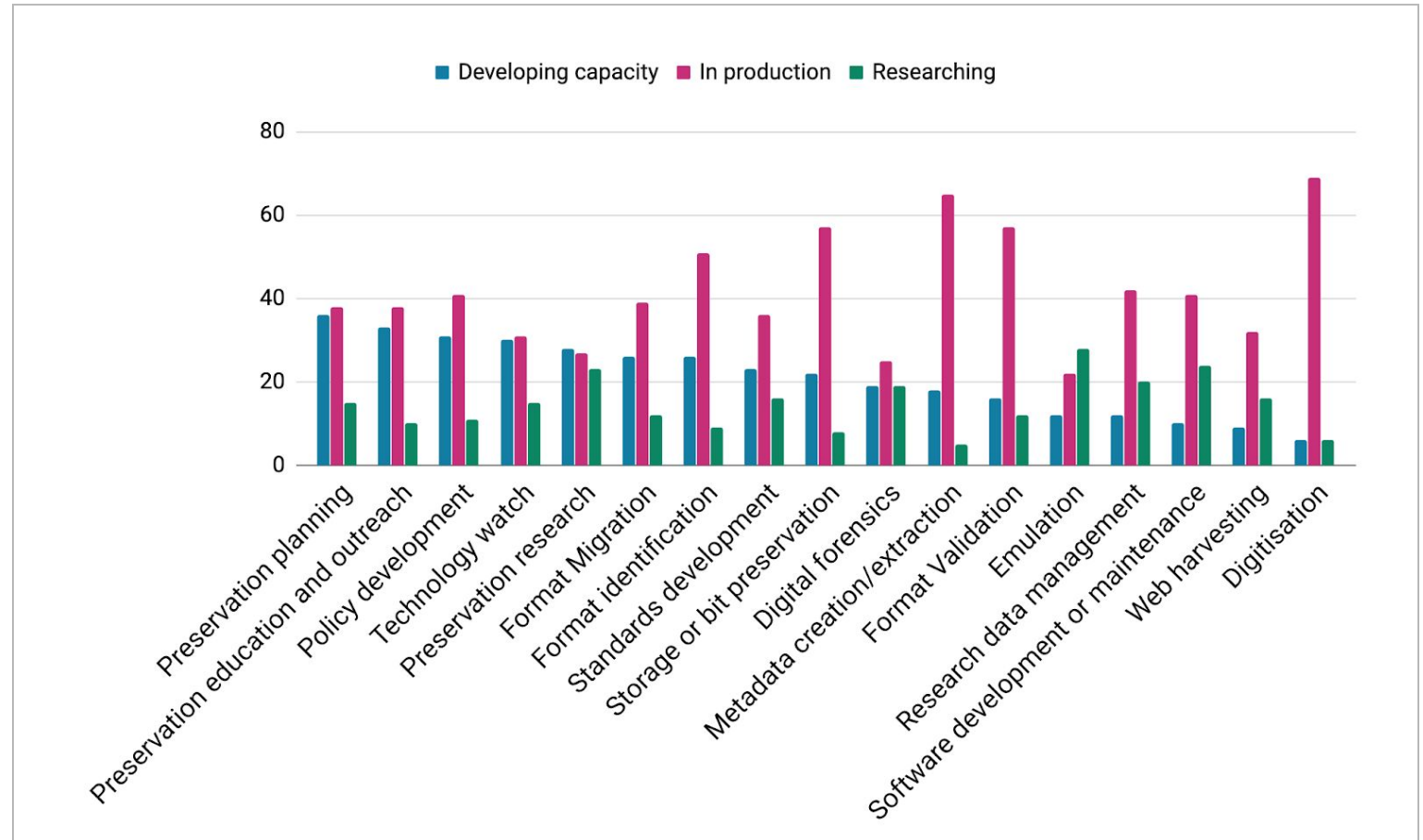

Fig. 10. Digital preservation activities by status

Some organisations are developing the capacity to carry out new activities. 36% are building capacity to carry out preservation planning, 33% for preservation education and outreach, and 31% for policy development. These organisations may be looking to set up digital preservation workflows for the first time, or perhaps undertaking a review of their existing policies and practices.

Despite research accounting for the highest average FTE (see fig. 5), there seems to be less capacity for research among respondents. Among the activities being researched, emulation and software development or maintenance are the most popular.

9

Digital Preservation Policies

Of those who responded to our question about policies (fig. 10), almost half (37 organisations) have a digital preservation policy, and 22% have a policy which is published openly. 20 organisations (26%) have a policy which is only available internally.

32% of organisations are currently developing a policy, and the remaining 19% neither have nor are developing a digital preservation policy.

Fig. 11. Organisations with digital preservation policies (n=77)
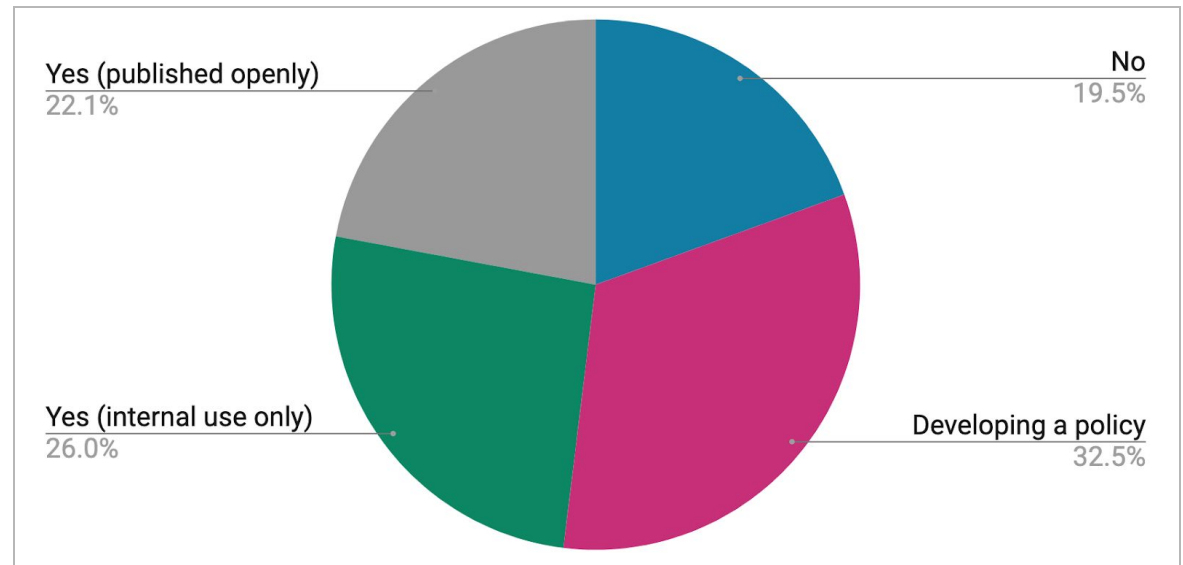
Type of organisation	Responses
Academic/research library	41
National library	6
National archive	6
Museum	9
Institutional archive	9
Service or infrastructure provider	7
Regional archive	7
Art Gallery	5
Software product or solution provider	3
Research unit	4
Funding body	1
State archive	2
Institutional library	1
Data centre	3
Regional library	3
Representative body	1
Municipality archive	1
Funding body	1
Commercial scientific archive	1

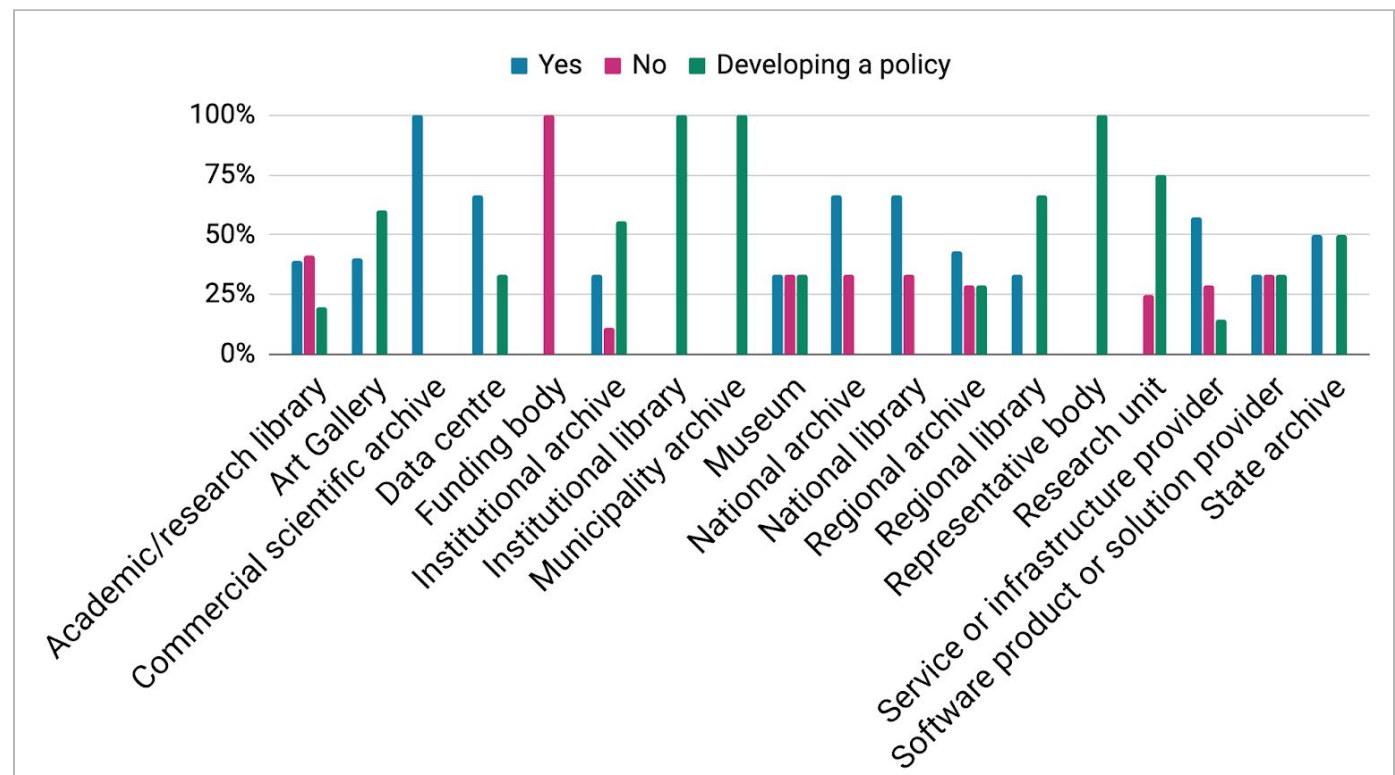*Fig. 12. Respondents to Q9 by type of organisation (n=77)*

Fig. 13. Digital preservation policies by type of organisation (n=77)

Approaches to creating a digital preservation policy...

'[Our] digital preservation policies are geared towards meeting the OAIS digital preservation model; our local policies are geared towards communicating these standards to a non-specialist audience of archivists.'

'We see our digital preservation policy as articulating a broad rationale and overarching expectations and commitments [...] We also intend to develop a digital preservation *plan* which contains specific actions and steps and is designed to carry out the *policy*.'

'Digital preservation is recognized as a core organizational function which is essential to [our] ability to support current and future research, teaching, and learning activities. We have had a preservation policy in some form for more than 10 years; it is reviewed and/or updated every two years.'

'We re-think our goals in preservation very regularly and adjust our initiative accordingly. We renew our published policy only on an irregular basis.'

'Our digital preservation policy details how we safeguard digitised and born-digital data over time, with the archiving of sufficient metadata and software tools to enable future access. Our DPP is guided by 7 main principles; Effective Governance, Sustainability, Transparency, Knowledge Transfer, Safeguarding Content, Access, and Third Party Alignment.'

'We normally take digitizing decisions as the cases arrive, but because the works are similar in nature and procedure so far, the same decisions and unwritten policies guide our digitization activities.'

'Although we have separate preservation policies for physical and for digital preservation, we plan to prepare an integrated approach to preservation of all library materials.'

Barriers to creating a digital preservation policy...

'The creation of a policy is complicated by the fact that we have multiple programs and each program has different priorities and expectations for longevity and access.'

'Because we work with many different institutional departments with different needs, our work in developing a policy about what we preserve and to what degree has been something of a moving target...'

'The major barrier to developing and implementing a digital preservation policy is time and resources, as well as the challenges of integrating digital preservation workflows into long standing traditional library and archival work done by staff who have little or no knowledge or experience in dealing with the issues related to the management of these formats.'

'The absence of a digital archivist has hampered the process and the fact that digital preservation is not currently considered to be a top level institutional priority at [our organisation].'

'Barriers: approval process for policies., multiple stakeholders involved and lack of clear ownership.'

'For us it's mostly about human resources (we need more humans). A digital preservation policy is on our radar but not a priority at this time.'

'The problem is [we] don't have enough time to write a policy, It's not yet as important as it should be for the administrators.'

Spend

In order to identify trends within the community, we asked respondents whether they track their spending on digital preservation activities, how they perceive their spending to have changed in the past five years, and how they believe their spending will change in the next five years.

Over half of organisations (around 51%) track their spending, while 36% do not (fig. 13).

Tracking spend appears to be more common in larger organisations, with 100% of those with 751-1000 members of staff answering yes, along with 71% of those with 501-750 and 65% of those with 201-500. However, 56% of very large organisations (those with more than 1000 members of staff) do not track their spending on digital preservation activities.
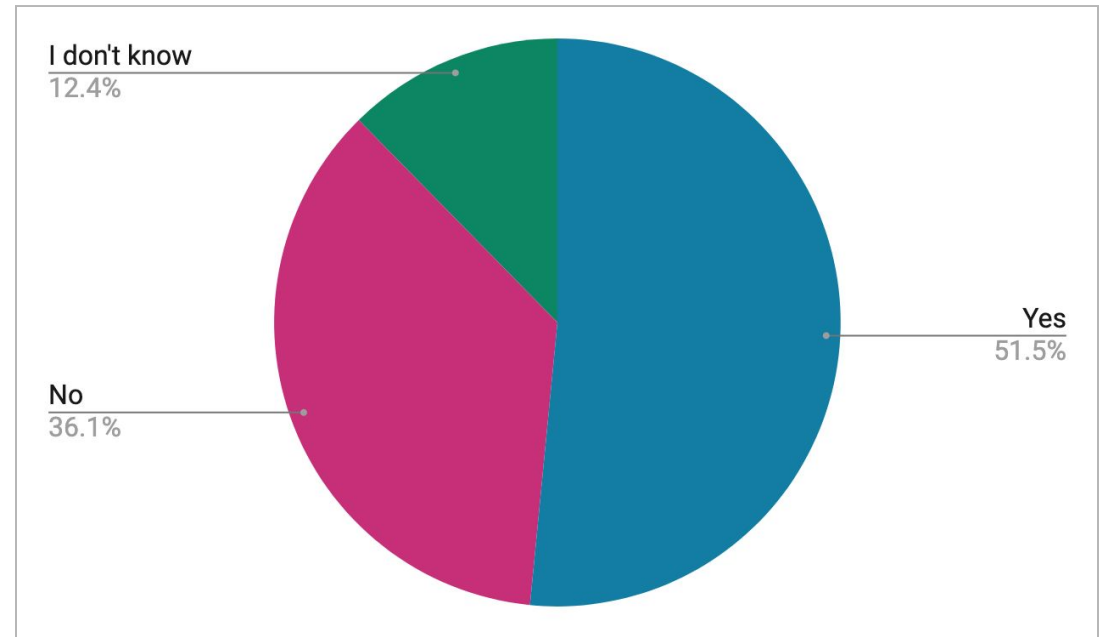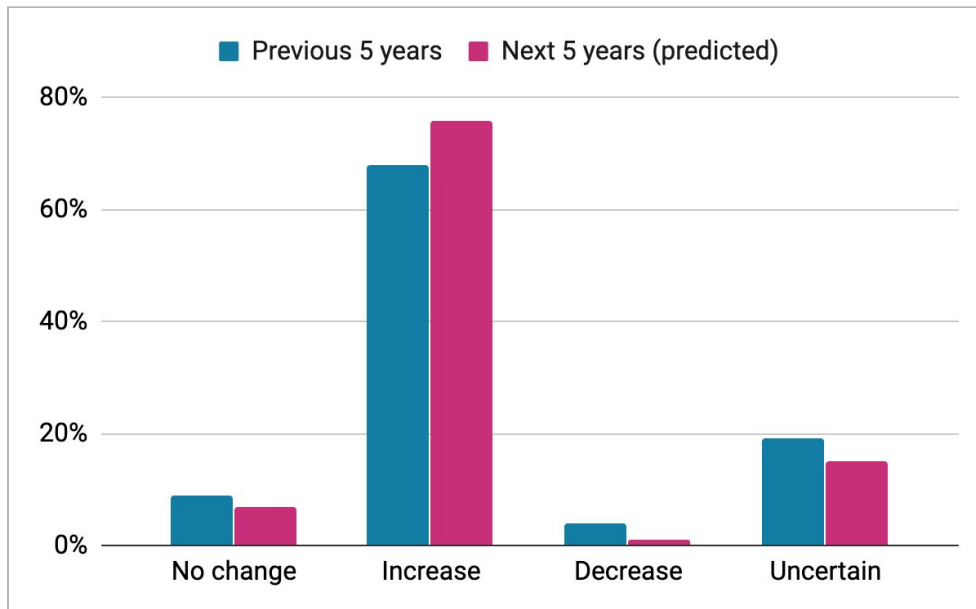

Fig. 14. Percentage of respondents who track their spend

74 respondents (68%) reported an increase in their digital preservation spending over the past five years. 9% believe that their spending has remained the same and 4% indicated a decrease in spending. 19% were uncertain.

76% of respondents predict the cost of their digital preservation activities to increase in the next five years. 7% expect no change in their spending. Just 1% believe their spending on digital preservation activities will decrease. 25% are uncertain.
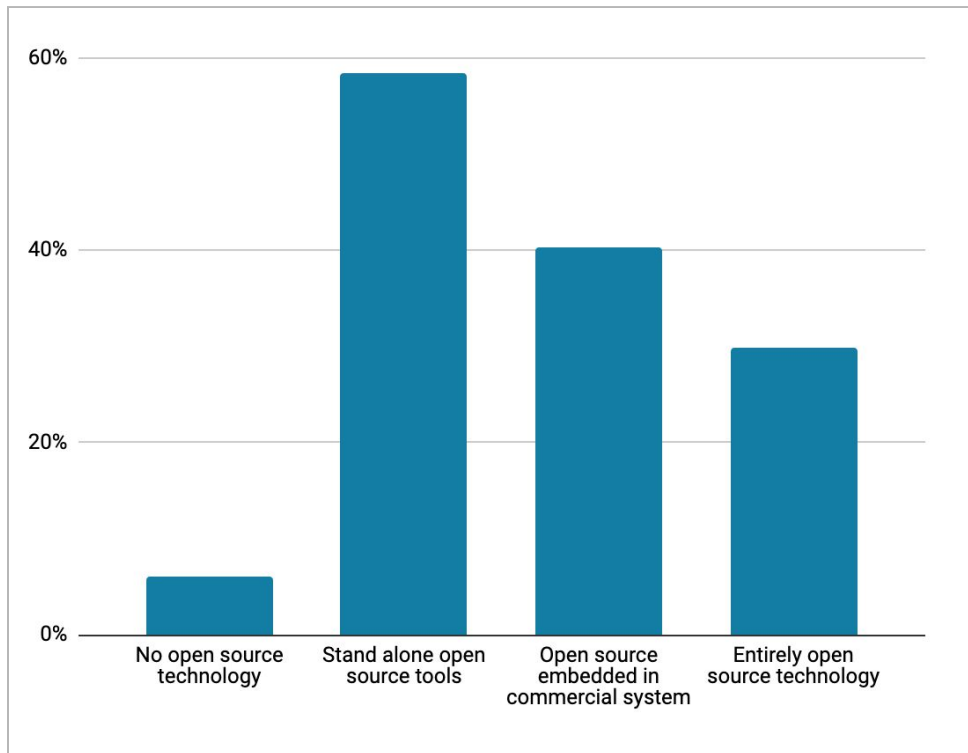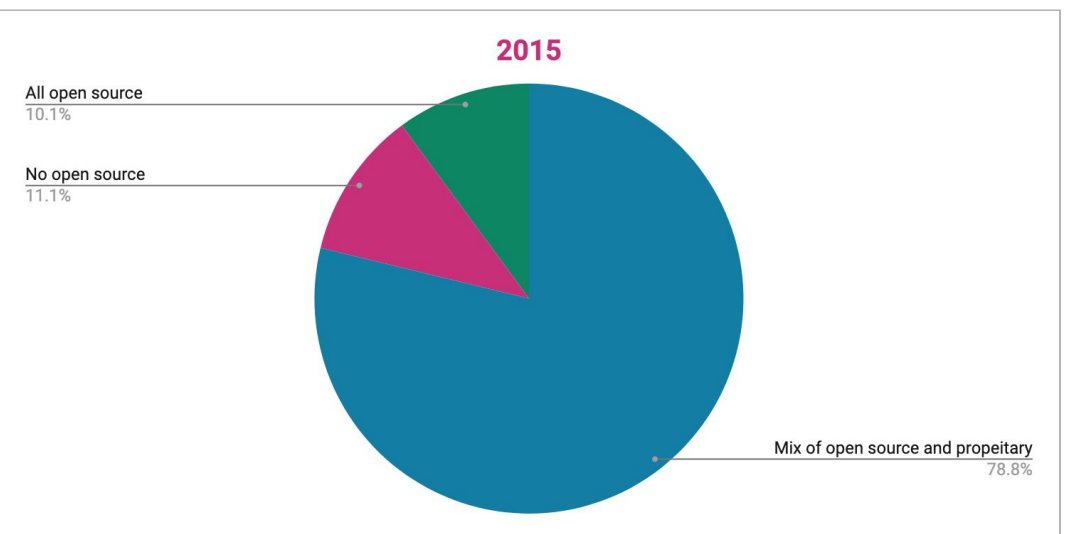
Fig. 15. Change in spending (n=97)

Open Source

Fig. 16. Use of open source technology (n=77)

Respondents in fact use proprietary digital preservation technology for a range of activities, including:

- Cataloguing and storage
- Digitisation and conversion
- Metadata embedding and extraction
- Virtualization
- Format identification and characterisation

We wanted to gain an understanding of the use of open source technologies across the community. 94% of respondents use some form of open source technology, with 68% using stand-alone open source tools (such as veraPDF and JHOVE), 40% using open source tools that are embedded in a commercial system.

Overall, the number of organisations using some form of open source has increased since 2015, when 11% of respondents said that they did not use any such technology, compared with just 6% today. According to a 2019 survey of our members, 100% of OPF member organisations use open source in one way or another.

30% of organisations use entirely open source technology, an increase from just 10% in 2015. This exceeds the expectations of respondents to the 2009 PLANETS survey, who predicted that 14% of organisations would use entirely open source software in the future. This group also anticipated that just 2% of future digital preservation workflows would utilise entirely proprietary solutions. As they expected, the majority now use a combination of open source and proprietary software.

Fig. 17. Use of open source technology in 2015

To gauge the level of familiarity with open source development practices among respondents, we asked about their participation in collaborative open source projects. 15% of organisations maintain or lead these kinds of projects, a decrease from 19% in 2015.

The number of respondents contributing to these kinds of projects has increased from 33% to 43%. 39% of respondents neither contribute to nor maintain open source collaborative digital preservation projects, down from 48% in 2015.

Engagement with open source project is higher among OPF membership, with 52% of organisations contributing and 33% maintaining or leading such initiatives.
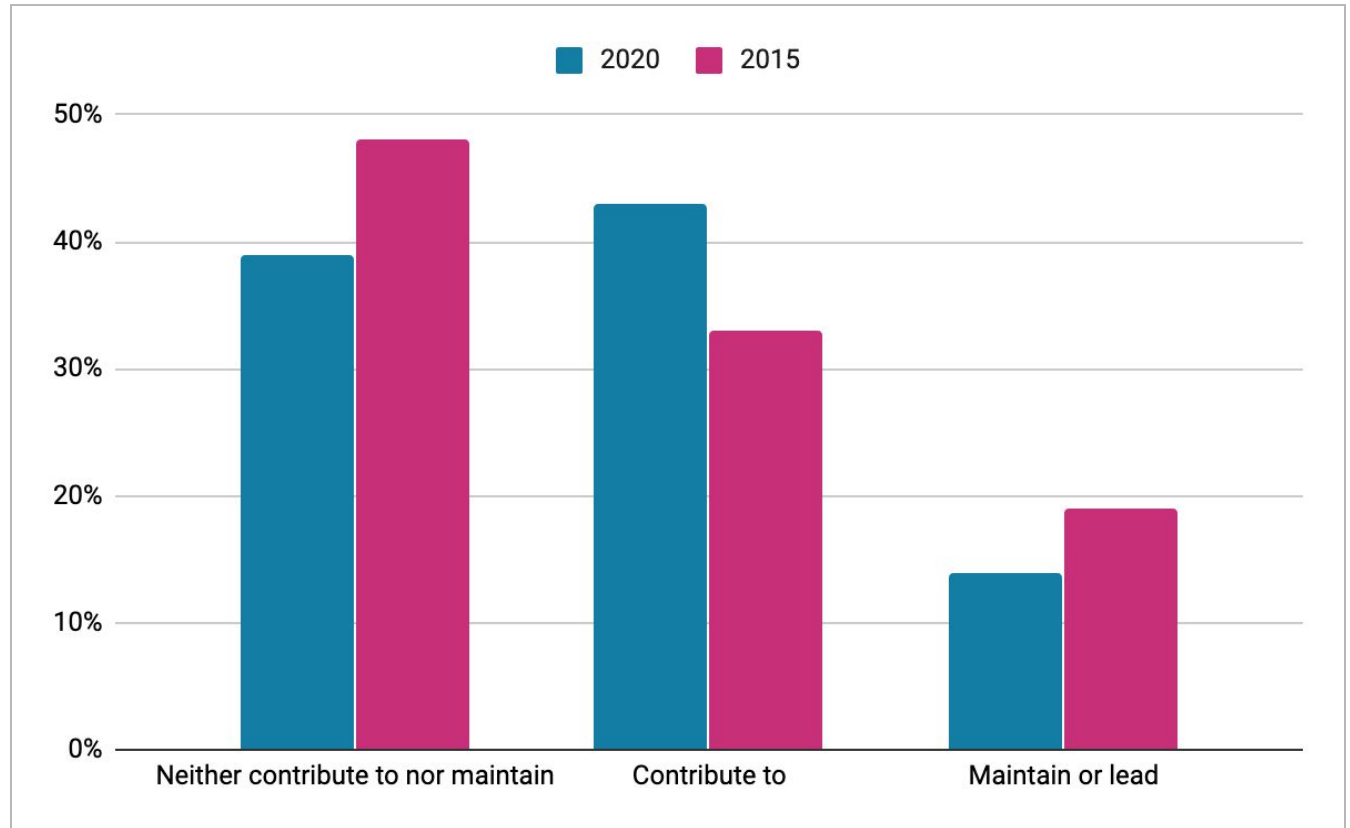

Fig. 18. Participation in collaborative open source projects, 2015 and 2020

Open source and proprietary technology...

'For us it's not a case of one or the other, it's a case of combining the two so we can create an enterpise-grade and scalable solution that follows community good practice and builds on community tools and solutions.'

'Open source benefits the entire community'

"Open source software is based on open standards, allowing integration with various other products. Although the cost, at the initial phase, may not necessarily be cheaper than a proprietary software.
Proprietary software requires users to agree to their licensing conditions. This is in contrast with open source which gives users the freedom to run the software for any purpose, on any number of machines etc."

'Multilingual support is often a major advantage to proprietary software. However this isn't always a limitation of open source software. Particularly as proprietary software moves toward subscription payment models, I am uncomfortable relying on it if I can avoid it. For most tasks, I find open source software meets our needs.'

'One of the biggest benefit of using open source software is setting up the collaborative environment...'

'We use proprietary software primarily due to a lack of staff resources to devote to open source software and an organizational/management bias against open source technologies. In addition, we have unique accessibility requirements and leadership felt it was better to try to hold a company to this, rather than develop this locally.

'The best part of open source software is customization according to our organization's requirements and also the cost benefit.'
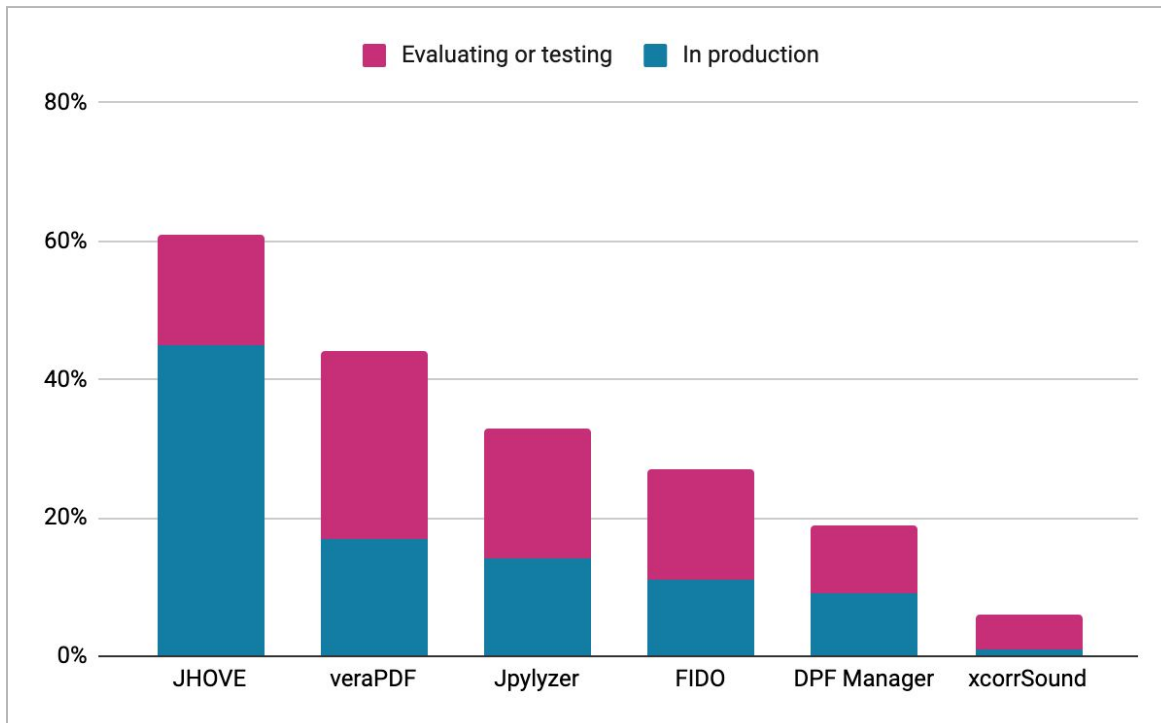
'The development and maintenance time to adapt open source software to our particular organizational needs can be extremely challenging, and in some ways the costs balance out due to this; however, we've generally found that the benefits of community and the ability to move our data when necessary without the risk of being trapped in a system we don't control when it no longer suits our needs outweigh the disadvantages.'

'Open source projects [...] tend to become obsolete when commitment and funding cease.'

'While open source software, when supported by the community is a huge asset, the staff resourcing costs to maintain open source within the organisation is not available to us. Reliance on others to fill the gaps would hamper us being a fully open source site.'

'We tend to use proprietary systems so that we have support from the vendor and we can ensure stability of the system as well as reduce the possibility of single point of failure or loss of institutional knowledge.'

Software

Fig. 19. Use of the OPF reference toolset

The OPF maintains a set of open source digital preservation tools: focussing collaborative effort and directing resources into effective solutions to common issues faced by memory organisations.

The tools that we currently maintain cover the pre-ingest and ingest stages of preservation, but our toolset is constantly evolving in response to the needs of our members.

Our products form a reference toolset for digital preservation which can be adapted for use in different organisational workflows.

In our 2015 community survey, JHOVE was used by 23% of respondents and was rated highest in terms of importance alongside DROID . These results contributed to our decision to adopt the software. Today, JHOVE remains the most widely used digital preservation tool and is used by 62% of organisations overall, with 45% using the tool as part of their production workflow and 16% evaluating or testing it. veraPDF is used by 45% of respondents and is the most commonly tested of our tools. 27 organisations (almost 30% of respondents) are testing it, while 17% are using it in production. Use of Fido in respondents' organisations is up from 10% in 2015 to 28% today, while Jpylyzer has seen a similar increase in popularity from 11% to 34%. DPF manager and xcorrSound are used by 19% and 6% of respondents respectively.

We asked about our respondents' use of other common tools, which were listed and described in the [survey appendix](#) (fig. 17).
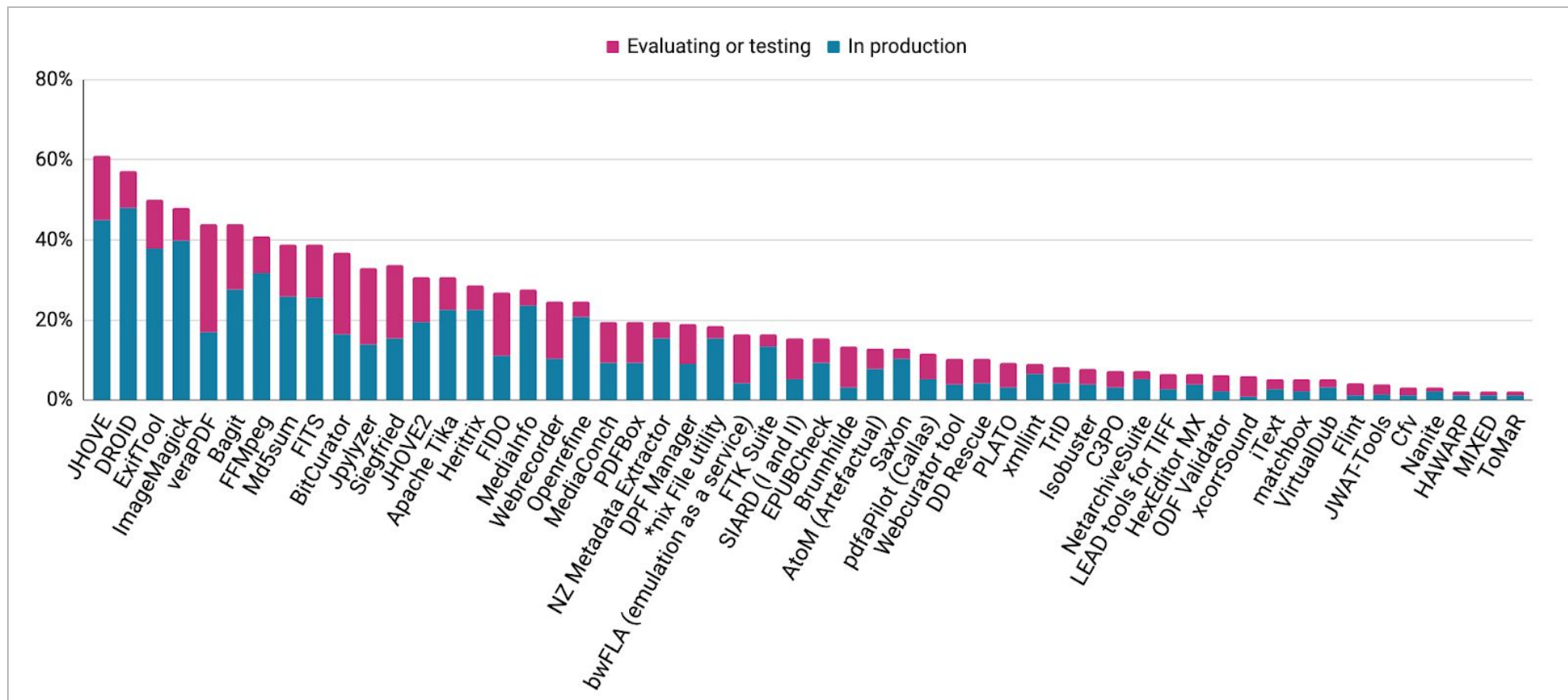

Fig. 20. Use of digital preservation tools

Second in popularity only to JHOVE, DROID is in use by 57% of organisations, up from 23% in 2015. Use of ExifTool has grown more widespread, from 27% in 2015 to 50% today, making it the third most popular tool. ImageMagick also remains popular in respondents' organisations. BitCurator and Siegfried are the most commonly tested tools.

Outside of the list provided, respondents use tools including Rodyn, Krool and Filescraper. Several organisations are developing additional functionality around Samvera framework technology, and some are involved in developing open source software, including the OPF AIG Spreadsheet Complexity Analyser and an in-house email archiving system and format validation tools. Other in-house development projects include tools and scripts for deposit, tools for USB imaging workflows, modules for processing XML-metadata, emulation tools, and a tool for aggregating pre-ingest statistics for collections comparing DROID, Siegfried and JHOVE outputs.

Community Resources

We asked which community resources are used by respondents. A list of suggested resources was provided. The most widely used resource is the OPF blog, which is the only one used by 100% of respondents. Other top resources in terms of use include the Library of Congress Recommended Formats Statement, the OPF Wiki, and the PRONOM format registry.

Respondents were also asked to indicate how important these resources are to the work that they carry out by rating them from one to five (five being the most important). The most highly valued resource among respondents is the NDSA Levels of Preservation, despite being used by the lowest percentage of respondents overall (83%). PRONOM is a long-standing favourite, highly valued by 2020 as well as 2015 respondents. Also popular are OPF resources such as Digipres.org, the OPF wiki, Digipres Q&A, and the OPF format corpus.
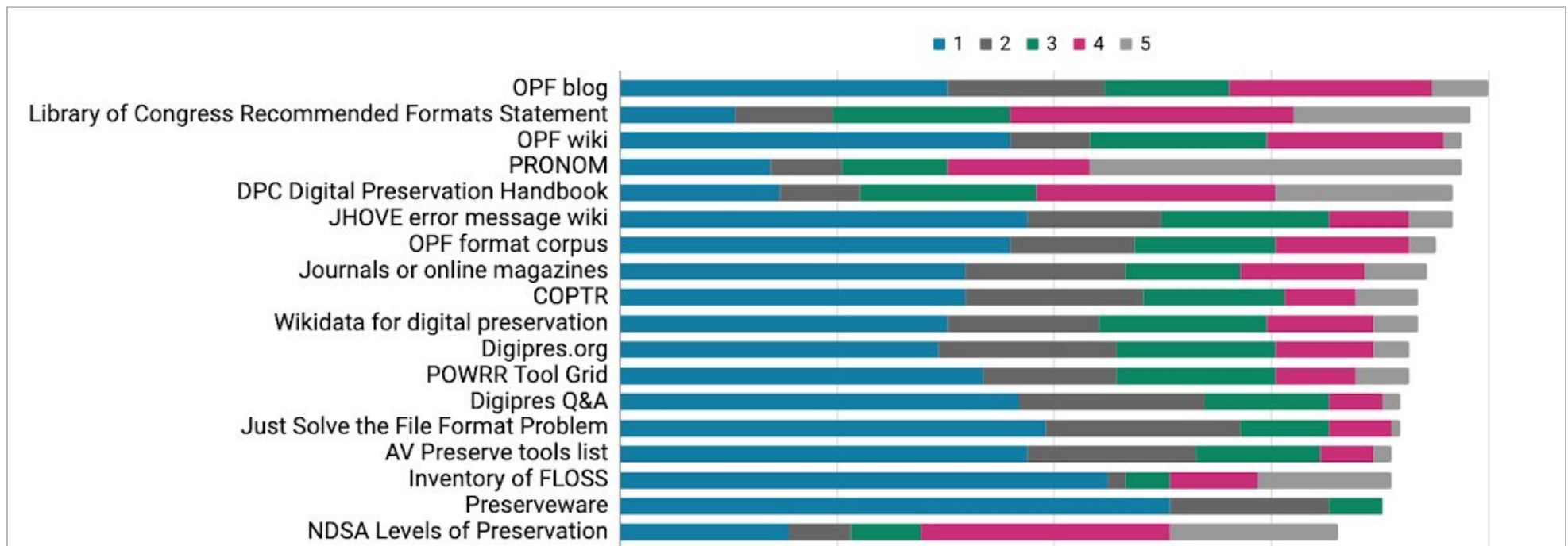

Fig. 21. Use and importance of community resources

Respondents were also asked to list any additional resources they use. The most popular resources included digital preservation mailing lists, twitter, proceedings from major digital preservation conferences (e.g. iPRES), and GitHub pages for tools such as JHOVE.

OPF knowledge sharing

Knowledge sharing is one of our key areas of strategic focus, and we facilitate a diverse range of activities to enable the dissemination of ideas, insights and good practices throughout the community.

The majority of respondents engage with our [community blog](#) in some way. The OPF blog features posts on a range of digital preservation topics including tools, project news, case studies and best practice. Posting on our blog is a great way to share your work and receive feedback from experts in a range of areas. Some respondents suggested that they would like to see more blog posts on preservation use cases.

> *OPF blogs are our primary international digital preservation communication channel, and helps us stay in touch with the international digital preservation community. Often these blogs contain practical solutions to daily digital preservation challenges.*
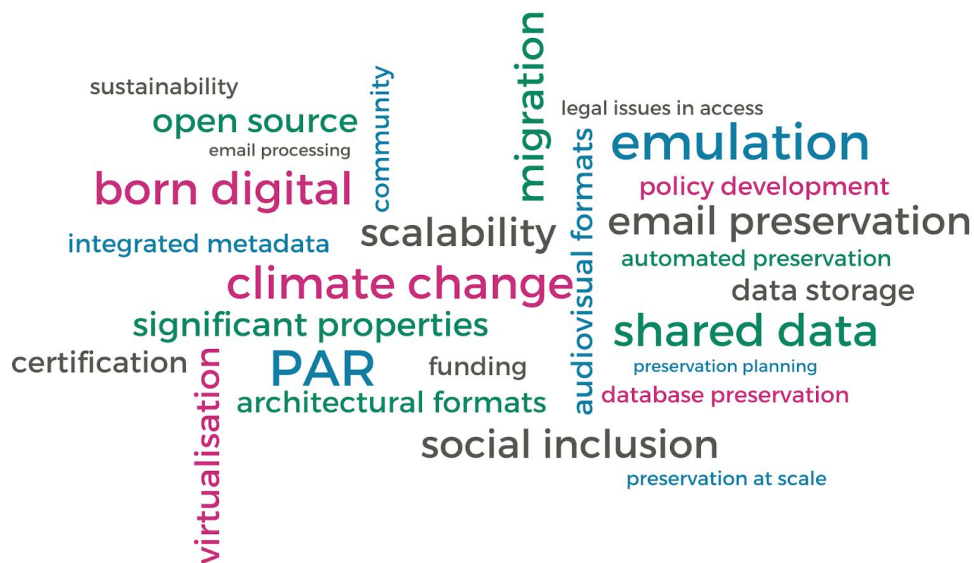
Since 2012, we've been running a popular webinar series that features expert speakers on a variety of topics at the forefront of digital preservation. Because these webinars are free to attend and open to everyone, they draw large audiences from organisations around the world. Of our 98 respondents, 52% have either given an OPF webinar, attended live, or watched a recording.

Finally we asked respondents to tell us which digital preservation 'hot topics' they would like to see included in our knowledge-sharing activities (fig. 19). Among the most popular topics were:

- shared data
- emulation
- climate change
- email preservation
- scalability

Fig. 22. Digital preservation hot topics

Conclusions

These results show us how organisations across the community approach their digital preservation work, what knowledge, tools and other resources they rely on to carry it out, and where the opportunities for further development lie. This information has enabled us to build a picture of the current state-of-play in digital preservation, which we will be reviewing in conjunction with our member survey and in comparison to other past surveys, identifying opportunities for growth as we continue working towards our vision of open, sustainable digital preservation.

Based on the results, we have identified several areas of strategic focus:

1. Tools

    The data from this survey provides us with opportunities to consider the composition of our reference toolset and set priorities for future development. We will refer to these results in our ongoing work to consolidate and enhance our reference toolset, ensuring that products like JHOVE, as the most widely used tool across the community, are fit for purpose for the long term. We will also be working to address any gaps in knowledge and expertise across the community and incorporate popular hot topics into our knowledge-sharing activities and resources.

    One such gap that we intend to explore is audiovisual content. Given the increasing prevalence of audiovisual content in collections across the community, we will work to ensure that popular audiovisual tools such as FFmpeg fit in naturally with the other products in our reference toolset.

2. Policy

    We recognise the community's need for guidance in developing digital preservation policies, and will consider activities, such as a working group, to respond to this need without duplicating existing efforts. Collaborative efforts are at the core of much of our work, and we believe that our expertise in sharing knowledge and best practice throughout the community can facilitate the removal of certain barriers to policy development. Given OPF's stewardship of the Preservation Action Registries initiative, we are particularly interested in finding ways to foster a better understanding of the relationship between basic preservation actions, tools, and automated procedures.

3. Open Source

    While the results of the survey show that more organisations are now using open source solutions, some misconceptions about open source practice and principles still exist. As leaders in open source development, we will continue our efforts to support users to learn about how open standards and solutions can empower organisations, drive progress, and create opportunities for collective growth.

Open
Preservation
Foundation

We would like to thank everyone who took part in the 2020 digital preservation community survey!

We are also grateful for the support of our members, whose ongoing commitment to OPF makes community initiatives like this survey possible.

WWW.OPENPRESERVATION.ORG

@OPENPRESERVE