# Preliminary Report

Web archiving project 2016

## Summary

This report shows the initial findings of a survey undertaken between November 2015 and January 2016 on the topic of web archiving with a specific focus on selection, description and access.

Those who responded to the survey self-identified as being part of web archiving processes within an organization. Respondents who provided country information showed that they were working in 13 different countries across five continents.

The findings of the report indicate that more research needs to be done, particularly in the areas of access and use of web archives, as well as who and how decisions are made in relation to recordkeeping processes over time.

Interviews will be conducted with participants interested in exploring process in depth between March and June 2016.

# Introduction

Web archiving survey launched November 15, 2015 and open till January 31, 2016.

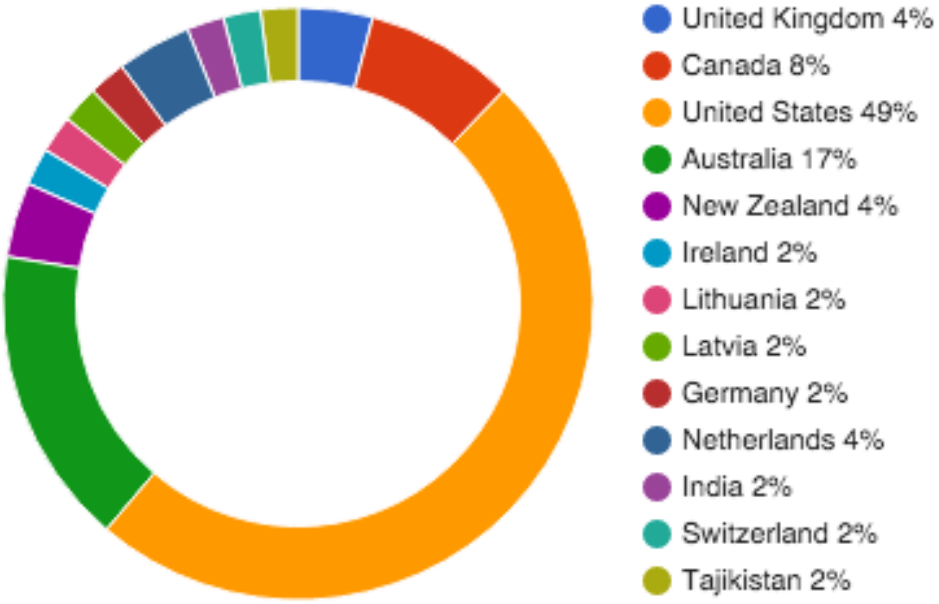
Purpose is to explore the practice and theory of web archiving to identify gaps and opportunities in relation to selection, description and access to web archives.

The data presented in this brief report will be presented to interviewees for discussion and will be made openly accessible at some point in 2016.

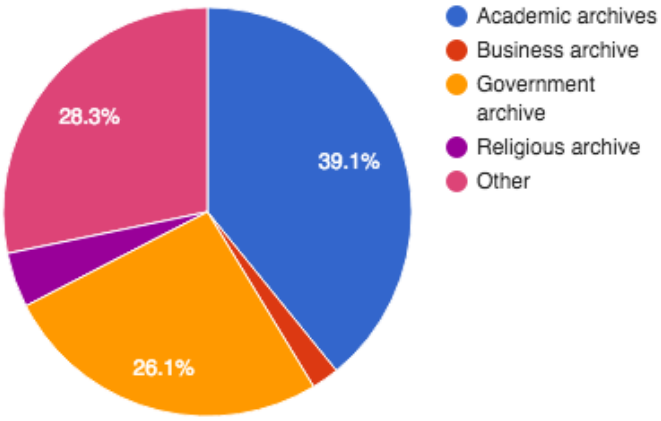
# Demographics

Total participants = 79
Participants who completed the survey = 54
Participants who were deemed ineligible (rejected) = 14
Participants who declined to continue = 1

### *Countries represented*



### *Institutions represented*




Other:

| | | | | |
|---|---|---|---|---|
| Academic library | Higher Education - Records Management | Library | National archive | National library |
| Academic University Library | Gallery Research Library | Museum library | National Cultural Organisation | Non-profit digital library and archive |

# Web archiving

### *Definition*

***"Web archiving is the process of collecting websites and the information that they contain and preserving them in an archive"***[1]

Agreed with provided definition = 60[2]
Did not agree = 3

Other terms and ideas offered in alternatives included other actions and content not mentioned in this definition:

- planning
- identification (selection, appraisal and acquisition)
- access (accessibility, use and rights)
- maintenance (ongoing preservation)
- metadata (capture and management of)

[1] Definition taken from Section 2.1.1 in *Web Archiving Guidance* by The National Archive (UK) p. 5. https://nationalarchives.gov.uk/documents/information-management/web-archiving-guidance.pdf
[2] Note that this question was before the qualifying question. So those that responded to this question may have been rejected from participation in the next question.

A participant provided a definition from the IIPC (International Internet Preservation Consortium)

***"The process of gathering up (harvesting) data that has been published on the Web, storing it, ensuring the data is preserved in an archive, and making the collected data available for future research."***[3]

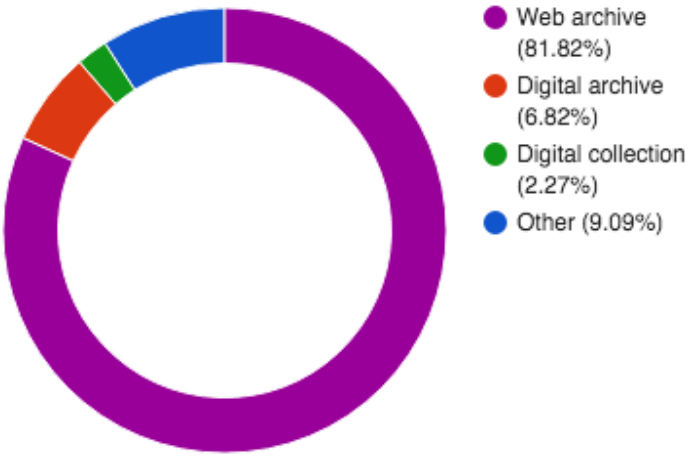
Not mentioned was some of the elements included in the Archive-It Web Archiving Life Cycle Model[4] (WALCM) that includes reference to:

- Policy
- Scoping
- Quality assurance and analysis
- Risk assessment
- Description and organization
- Workflow
- Reuse
- Planning, vision and objectives

While the WALCM is a workflow model, not a definition, some of the participants made it clear that elements related to workflow were important to include in the definition. Most who offered alternatives identified the steps of capture, preservation and access (in that order, although these exact words may not have been used). Others included reference to a "digital life cycle" including planning, capture of changes to websites over time, maintenance, and facilitation of access.

# Identity

***How does your organization describe the web archive?***
Respondents = 44




Other responses:

[3] para. 1. http://netpreserve.org/about-us
[4] p. 3. http://ait.blog.archive.org/files/2014/04/archiveit_life_cycle_model.pdf

- Not described
- E-records taken from the university's website.
- Intranet Archive
- Web Site Captures

## *Purpose of web archive*

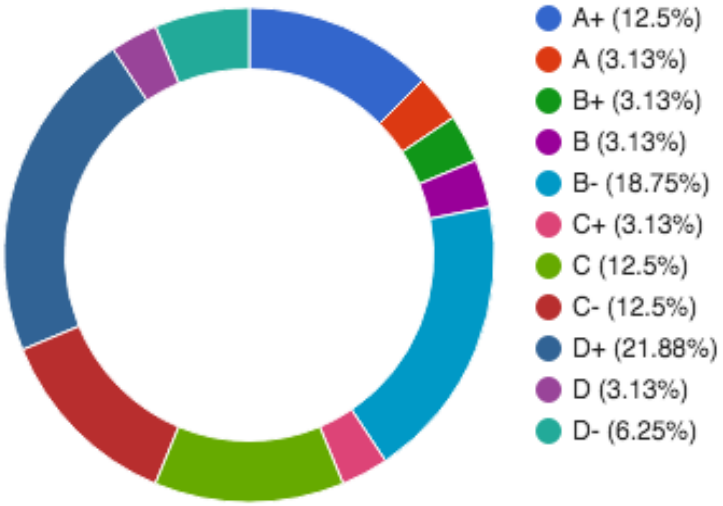
There was a question related to how the archive was conceived which is linked to the potential mission or vision of the archive, as well as how it might be carried out. Respondents (44) were asked to identify the primary driver for conceptualization from a list that included event, format, domain, organization or individual or content creator driven. This question forced respondents to answer only one option.

While 44 respondents answered this question and the two main drivers were domain and organization (27.7% each), there were 11 respondents (25%) who indicated “other”. In the “other” category many respondents indicated that there was no one primary driver but it was more a combination or the drivers were different for different collections. One respondent linked this process to the term acquisition.

## *Self-identified grade*

Respondents = 32
Missing grades = A- and F (0%)

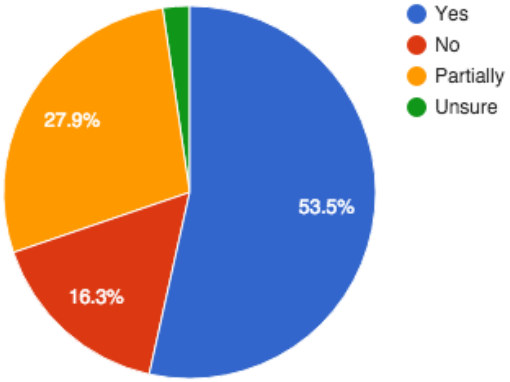
Note that question did not ask for success or failure of any particular part or practice, but asked for a grade on the organization’s web archiving. Therefore this grade could refer to the organization itself as supporter/facilitator and/or any element or part of the web archiving process and practices.



# Framework

***Is your organization's web archiving supported by an official web archiving program consisting of policy, processes, people and technology?***

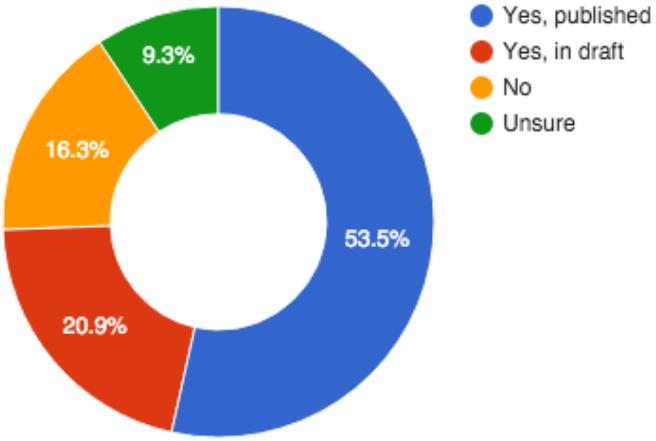
Respondents = 43




There was an additional question asking if the policies and processes in the organization were useful across a scale of 1(not at all useful) to 5 (extremely useful). There were only 23 respondents to this question and the findings suggest that perceived usefulness policies and processes were close to the mean (3).

## *Formal collections policy*

Respondents = 43

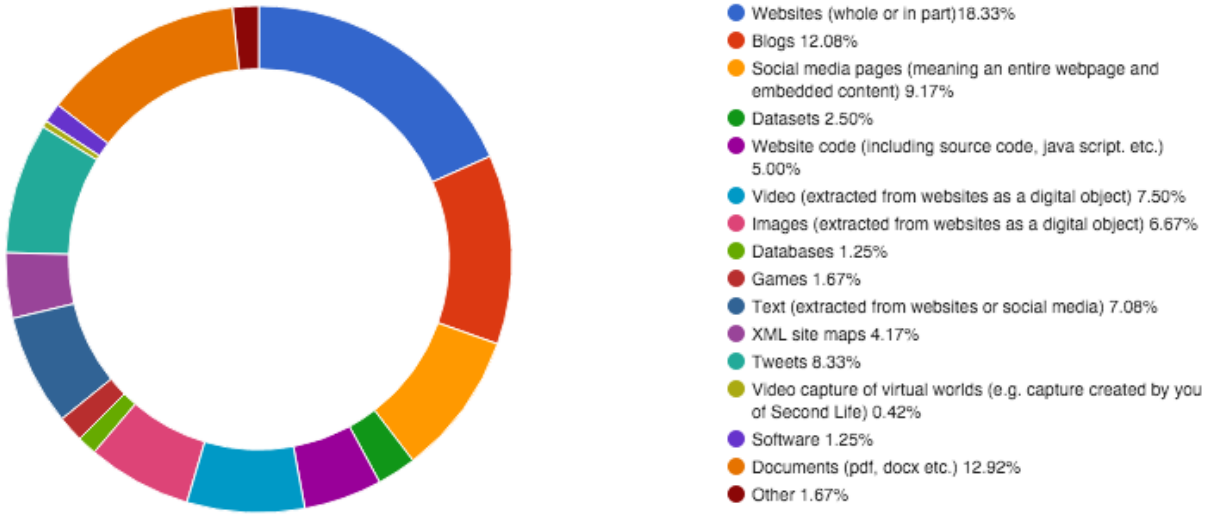



There was also another question about policy related to whether or not web content was specifically mentioned. There is a relationship between these questions where the first asks if there is a formal collections policy and the second asks if web content is mentioned in the formal collections policy. The second question responses (34 respondents) indicated that in

44.1% of responses there was no mention of web content or the respondents were unsure. This is an area for further questioning including why respondents might not know what is in the collections policy and why the policy, which was applied to this context, might not include reference to web content. It is possible that a collections or similar kind of policy may address requirements rather than formats and could be nonspecific to format. This requires additional investigation.

## *Overview of content collected*

Responses = 240 (this was a select all that apply question)




Summarized other category:

- 'whole website' capture will often include javascript, documents, video, etc.
- Digital serials, such as newsletters and other ephemera
- any content on the web is in scope

## *Metadata*

Questions about metadata included what types were collected, what was done with the harvested metadata and what standards or schemas were used. The first question was answered by 44 respondents, the second was answered by only 39 and the third by 50 respondents.

In addition to the choices given, respondents clarified and added detail. In particular, it is noted that the responses here raise questions about **when** metadata is managed, not just how. Managed refers not just to collection, but identification and use in the process including in finding aids and in the preservation system (implications for workflow/cycle).

| | | |
|---|---|---|
| Metadata is not collected | 5 | 11.36% |
| Technical metadata only is collected | 1 | 2.27% |
| Descriptive metadata only is collected | 9 | 20.45% |
| Technical and descriptive metadata is collected | 15 | 34.09% |
| Unsure | 9 | 20.45% |
| Other | 5 | 11.36% |

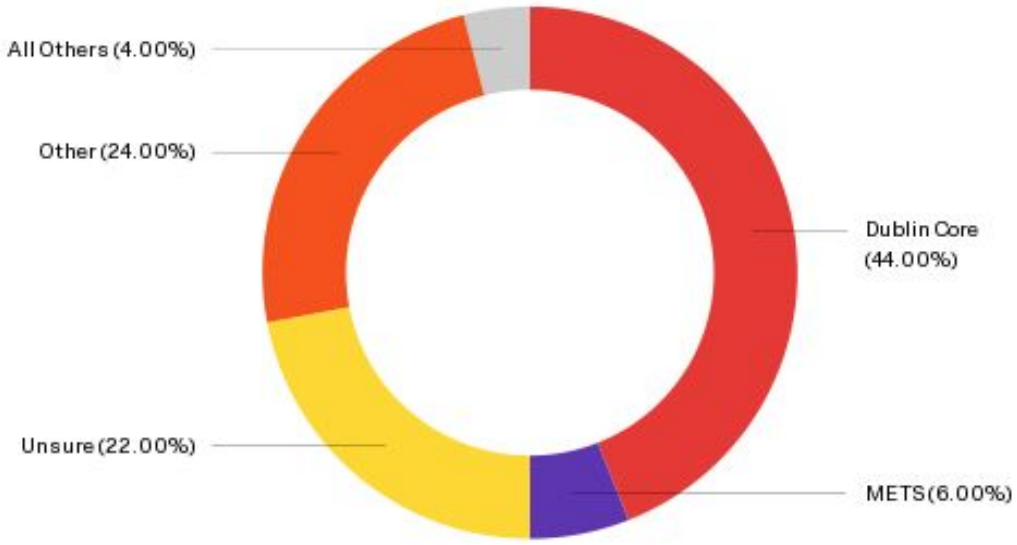
Other:

- all of the above
- whatever is in the WARC and archival finding aid description
- Some preservation metadata is collected.
- Descriptive metadata is applied rather than collected. Technical metadata may be captured, but I have not viewed the WARC files.
- We are just starting and have not yet finalized metadata

| | | |
|---|---|---|
| Any relevant information is added to the archive metadata records | 14 | 35.90% |
| It is included as part of the digital object and is preserved | 7 | 17.95% |
| Any relevant text is added to a finding aid | 2 | 5.13% |
| All of the above | 4 | 10.26% |
| None of the above | 1 | 2.56% |
| Unsure | 9 | 23.08% |
| Other | 2 | 5.13% |

Other:

- Added to Archive-It interface; MARC records also added to our OPAC.
- Added and included; but not in "finding aids"




Other metadata standards identified by respondents:

- Heretrix
- AACR2
- LCSH
- DDC
- MARC21
- Dublin Core

Other answers included no use of any standard, or lack of identification of a standard, as well as planned use of a standard. Standards identified as being planned for use were:

- Metadata Technical Specification associated with the Records Management Standard
- METS
- LMER

Also note that PREMIS data dictionary was not identified as being used by any respondents.

Respondents (37) responded to whether or not urls for current (active) websites were being added to the metadata as part of the archiving processes. Of these 37, 4 indicated current urls were not added and 11 were not sure.

## *Permissions*

Respondents = 44

| | | |
|---|---|---|
| Permission is not obtained as the content is public domain | 7 | 15.91% |

| | | |
|---|---|---|
| Collection of content is authorized via legislation or other legal mandate | 5 | 11.36% |
| Copyright of the web content is owned by the organization collecting it | 9 | 20.45% |
| Working with content creators to negotiate permission | 5 | 11.36% |
| Web content is donated and the donor has provided permission | 3 | 6.82% |
| Permission is not obtained | 3 | 6.82% |
| Unsure | 1 | 2.27% |
| Other | 11 | 25.00% |

Note that that the first question including “public domain” may be seen as provocative as it may not be entirely clear how this is defined. This question also has significant jurisdictional implications based on location and legal mandates.

Other:

- Answer four via opt-out procedure
- Only archive content we create
- we have a mix of public domain and sites we have to notify or obtain permission from
- Public domain has a specific meaning that is not quite correct. I would say that "Permission is not obtained as the content is public" but I don't know if that is the same as "Permission is not obtained" - feel free to recode as appropriate.
- We do not seek permission for public domain materials but notify private individuals/organizations of our activity, giving them the option to opt out.
- Either authorised by legal deposit or permission obtained if it's outside legal deposit (i.e. published overseas)
- We notify but rarely request permission multiple methods listed
- In some cases we give notice and in others we ask permission, based on a fair use analysis.
- Copyright of the web content is owned by the organization collecting it for university archives materials. Web content is donated for manuscript

## *Access*

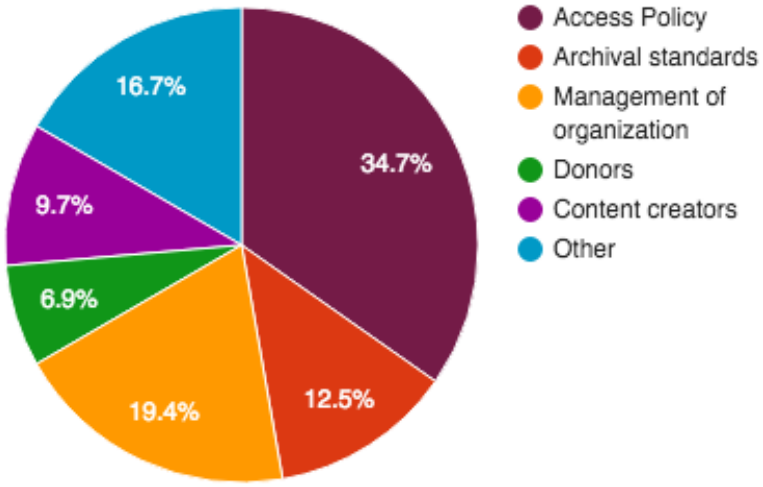
Questions about access included how access conditions were decided on, how accessible the web archive was, and whether or not information about access was being recorded and what it is used for. Note that questions about access, accessibility, and assessment or evaluation of these can be interpreted in various ways.

The accessibility question showed much more variance than previous questions using a scale. Respondents (30) indicated if the contents of their archive was between 1 (not accessible at all) and 5 (highly accessible). The mean of 3.33 with a standard deviation of 1.58 and variance of

2.49 indicated there were several answers at the extremes of the scale. However, overall the mean indicates many identified their archive as accessible.

The responses (72) related to how decisions were made showed a mix of influences, as expected. There responses in the “other” section showed that some web archives have significant and highly restrictive access based on organizational needs. Two additional questions related to identification of users who access and the purpose for accessing the web archives indicated that information is not being captured and/or evaluation is not being undertaken to identify those accessing the web archives and for what purpose they are using it for.

There needs to be some identification and exploration of the role and identity of organizational archives as cultural heritage. Are they/aren’t they?




Other:

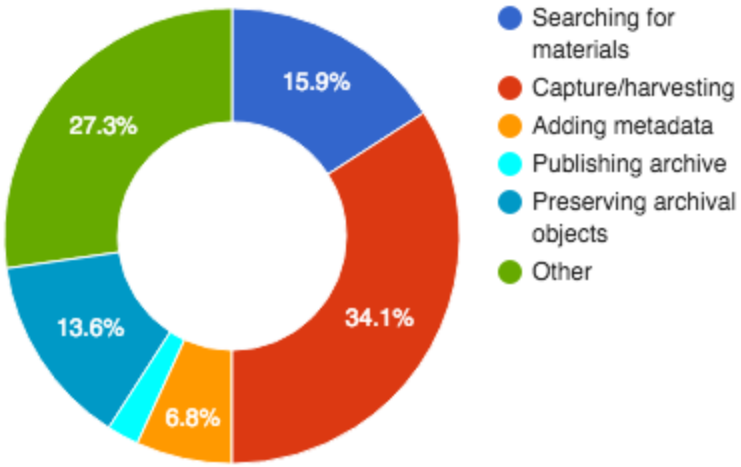
- Legal issues (especially the Copyright Act)
- Only IT can access
- Our office of general counsel provides guidance to us
- Law (legal deposit), access only within the reading rooms
- Combination of legislation (eg privacy, public records), standards and internal access rules.
- No access conditions apply. Only the website author/administrator has access.
- National cultural organisation with legal mandate to make collections accessible to public
- Legal deposit legislation
- unwritten access policy:)
- Access is open by default; however, this is not a formal policy
- The web archive is publicly available

- subjective judgment of the digital archivist. The default is to provide public access. There has been only one exception thus far.

# Technologies

There were two questions specific to technologies. The first asked about suitability related to the web archiving processes. The respondents (34) were asked to use a sliding scale from 1 (unsuitable) to 5 (extremely suitable). The responses indicate that most people measured suitability around the 3 (mean) with variation up to approx. 1.5 away from the mean.

The second question asked what processes could do with better technologies and respondents (44) were forced to choose only one answer which could provide some clue to ranking. However, what happened is a significant proportion respondents chose 'other' instead.




'Other' suggestions are summarized:

- all of the above
- quality control and assurance for various processes but also to check completeness
- curating web archive collections including publishing the archive
- automation and scalable tools including identification of obsolete documents and image files

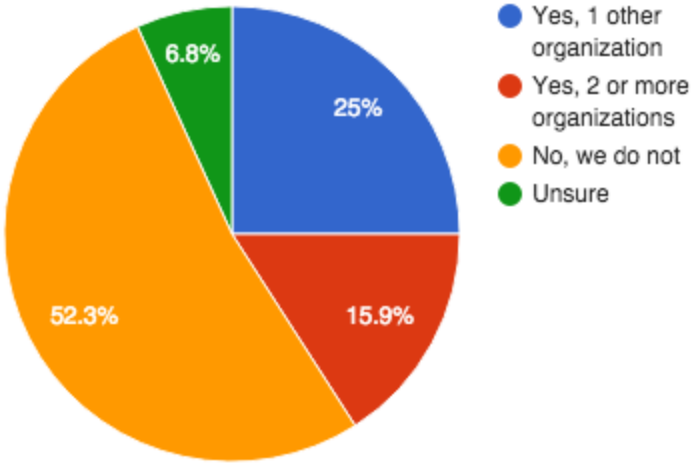
# Community interaction

These questions were added to enquire about how the identification and management of web archives might be undertaken in conjunction with those who might be considered the community. Community was used instead of stakeholders as this is the current term referenced by OAIS and DCC Lifecycle Models. How community is defined was left intentionally vague and is an aspect to be followed up in the interviewing process.

The first question represented below with the pie chart asked if the web archiving organization worked with other organizations. The second question asked if the organization made direct contact with individuals, community groups, or NGOs external to the organization that influence or impact on decisions related to what is collected, how it is described and/or how it can be accessed. This second question was answered by 43 respondents with 16 saying they did have contact, 20 saying they had no contact with anyone external to the organization and 7 were unsure. These answers do not necessarily indicate that there is a lack of interaction, but that potentially the community the web archiving is linked to may be internal to the organization. However, overall, there appears to be little contact related to web archiving outside of the organization. This is an area flagged for further investigation with particular reference to use of web archives.

*Working with other organizations*

Respondents = 44




# General findings

- Over 70% of those who responded to the survey question about length of time web archiving (44) indicated that their organization has had a web archive or has conducting web archiving for over 2 years.
- Of these nearly a third indicated their web archiving practices had been going for over 10 years.
- 27.3% indicated web archiving practices were less than 2 years old in their organization.
- The responses to the question asking who does best practice web archiving heralded a variety of answers including several “I do not knows” which are summarized below:
    - Library of Congress (USA)
    - University of Michigan (USA)
    - British Library x6 (UK)
    - Melbourne University (AUS)
    - UK Archives (unclear of relationship to British Library) (UK)
    - National Library of Australia (AUS)
    - New York Art Resources Consortium (USA)
    - Columbia University x2 (USA)
    - Internet Archive/Archive IT x 7 (USA)
    - Stanford University Libraries x 2 (USA)
    - Japanese National Library (JAP)
    - The Icelandic National Library (ICE)
- When asked if there were further topics or questions to be asked 15 people responded with varied answers, summarized and classified below:
    - Management, frameworks and general including:

        - Quality assurance
        - Program management
        - Setting up and making a case for web archiving
            - Advocacy
        - Scale
        - Collaborative archiving models
        - Strategies for adding value
        - Struggles generally
        - Complexities of different purposes of web archiving (organizational archive versus cultural heritage)
        - Membership of international consortiums
        - Standards (WARC mentioned specifically). No one else mentioned standards.
    - Processes including:
        - Catalogs and cataloging
        - Curation (related also to adding value)
        - Appraisal including retention and disposal
        - Who is responsible and what do they do
        - Frequency and completeness of capture
        - Acquisition
            - hidden web and intranets
        - Ongoing preservation strategies
            - Active or bit-level preservation of web archive data
    - Technologies
        - Use of different platforms and/or tools
        - Service providers
    - Access and use
        - User studies
        - Limitations on access/access controls
        - How accessed
        - Rights

# Analysis work still to be done

This report provides information on the face value of the responses. I have not reported on any cross-tabulations of data. This work is still in progress.

# Conclusions

While it is too early for conclusions it is clear there is much to be done in the world of web archiving. In particular, it seems essential to look at the different influences on web archiving

policies and practices. The survey did not mention much about appraisal, arrangement and description other than use of metadata schemas and standards and it was not really flagged as issues that required more investigation by the participants. Does this mean that how web archiving is conceptually understood as part of recordkeeping is matured and requires no action?

Aspects that stand out for further and more detailed research:

- What decisions are made and at what points in time. Does a lifecycle model represent all the decisions being made and their impact on web archiving? Further investigation into the actual processes of web archiving including ongoing practices in preservation and quality controls (and reappraisal) are part of this research. Defining where and how decision making happens and the role of decision making in archiving is important to this research. How policy is linked to decision making is also key to this research.
- How communities are conceptualized and included in decision making appears to be an area that should be investigated in more detail. This area of research could potentially address elements related to access - who needs it, how they need it and what people are doing with web archives and how this might influence web archiving practices and policies. It would also include how a community is defined and what it means to interact with communities that have different roles to play. There is some work done in this area, but it could be developed more.
- The relationship that OAIS and DCC Models have to web archiving. OAIS and DCC were created within research data and archives communities. Do they translate to web archiving that is being performed in the profession?

For further information or enquiries please contact the researcher:

**Leisa Gibbons Ph.D.**
Assistant Professor, School of Library and Information Science
lgibbon3@kent.edu
Direct: +1-330-672-0014
314 University Library
Kent, OH 44242-0001
www.kent.edu