# SPN Metadata Standards Survey:
# Initial Results, Analysis, and Next Steps

**SPN Metadata Standards and Policies Working Group:**
Glynn Edwards, Wendy Hagenmaier, Eric Kaltman, Daniel Noonan,
Elizabeth Roke, Katherine Thornton and Tim Walsh

**October 2017**

## Introduction

The Software Preservation Network (SPN) strives to solicit community input and to build consensus around software preservation strategies, as part of the larger effort to ensure long-term access to digital cultural heritage. The SPN's Metadata Standards and Policies Working Group (Working Group) is responsible for developing, promoting and advocating for common metadata frameworks and related metadata standards, vocabularies, and ontologies that support software preservation and access. To that end, the Working Group developed and deployed the *SPN Metadata Survey* in April of 2017 to provide our Working Group with a baseline of practice regarding metadata being used to describe and manage the preservation of software, and to identify where there may be gaps in practices and standards. There is an expectation that the survey results will inform not only our Working Group, but all the SPN working groups as we progress on developing best practices, guidelines and minimal requirements for software preservation.

### *Survey distribution*

The Working Group sent the survey to multiple electronic mailing lists and attempted to get greater exposure via SAA Electronic Records Section's Twitter feed. While we did reach most of the groups we listed as important, a few were missed, primarily due to lack of membership among our Working Group members.

### *Were there any in-depth interviews?*

The Working Group has not performed any in-depth interviews at this time, although 43% of the respondents stated they were willing to be interviewed. Committee members' time has been the main constraint to pursue this additional information.

### *Design and analysis*

The Working Group met over a few weeks in early 2017 to develop the questions based upon the desired data to be collected. Dan Noonan designed the survey utilizing Google Forms with modifications after a period of Working Group review. Working Group members posted invitations to participate in the survey to Twitter, as well as those electronic mailing lists to which they were subscribers. Noonan conducted the initial analysis and visualization of the survey data.

## *Selected Demographics*

### *Geographic*

Most of the 44 respondents were from institutions in the United States with 18% responding from the United Kingdom, European Union, New Zealand and the Netherlands. It was interesting that only one respondent from outside the U.S. was from an academic institution; the others were predominantly from government agencies, such as national archives and libraries.

### *Job Title*

Only 43% of the respondents provided job titles and this correlated to those who were willing to be interviewed. This might suggest that most of the institutions responding have either not yet determined whether they will collect software or which department or positions would be responsible.

Of those 19 respondents who did provide job titles, there appeared to be a dividing line between those that might be construed as falling in the digital preservation arena versus those that fall under the auspices of a collection services group. While the numbers are fairly even, the job titles and specific responsibilities are more nuanced.

Based on our informal distinctions between positions that fall on the two sides of the curation continuum, we found there was less variety for job titles under those we identified as collection services, most falling into a variation on “digital archivist.” Those that were on the digital preservation side covered much more ground. (See Appendix D for examples.)

### *Institutional unit*

The survey asked respondents to indicate what institutional unit to which they belonged. The majority of respondents (59.1%) stated they are located in either an archive and special collections unit or a library. The remainder were more difficult to aggregate as the units as named were hard to define. While multiple units (29.5%) dealt with preservation–collection care, preservation, digital archives, and digital preservation—it was unclear where these units are within their larger organizational context. The remaining 11.4% identified themselves as belonging to digital services, engineering, IT, and software and games.[1]

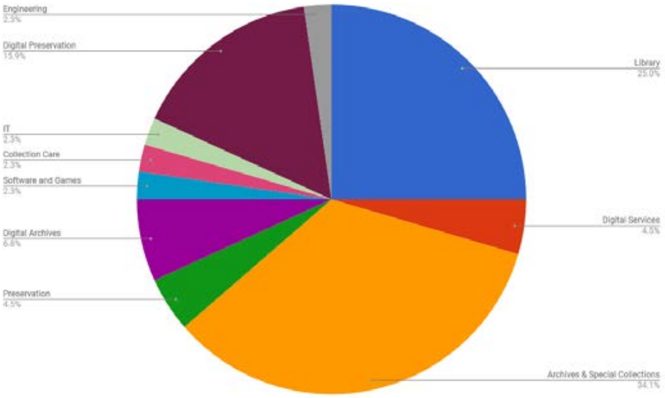


*Figure 1: Institutional Unit (Normalized)*

# Detailed Survey Analysis

## *Collection Development*

Most institutions are lagging in addressing the acquisition of software in their collecting policies with over 84% of the respondents stating that they had no specific language in their policies concerning software. Seven institutions replied that they did have a collection development policy that addressed collecting software. Unfortunately, we did not ask them to include a link if published online. Here are a few examples discovered via our own searches:

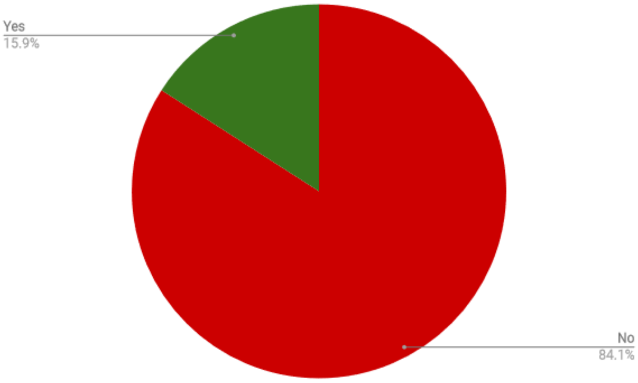


*Figure 2: Does your institution's collection development policy address collecting software?*

- MIT's Institute Archives & Special Collections does have an online policy regarding Digital Archives Practice in which they state "The Archives will not keep copies of commercial software, operating system software, nor other non-archival material that may be transferred to the Archives in the course of transferring digital material."
- The Computer History Museum has launched the Center for Software History with a mission statement focused on collecting software source code.
- Georgia Tech Archives also has a retroTECH Collection Development Policy available online that includes the acquisition of software. Their policy statement includes a list of "software of particular interest to retroTECH."

Other institutions that stated they did have collection development policies may keep these internal, as it was difficult to find them online.

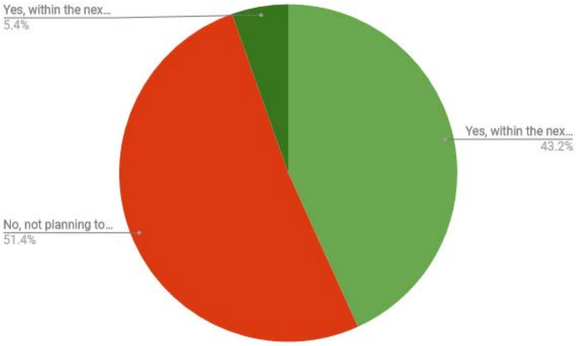


*Figure 3: For those who do not have a Collection Development Policy that addresses software, do you plan to update your institution's collection development policy to account for software?*

For those 37 repositories that did not address software in their collecting policies, over half (51.4%) had no plans to review their policies in the next few years. Of those who were planning to reevaluate their policies and include software, only a very small percentage (5.4%) had plans to do this in the upcoming year. 15 of the 17 respondents do not have immediate plans to reexamine policies toward software acquisition.

## *Donor agreements*

Of the 44 respondents, only eight said their donor agreements or accessioning documents address collecting software. For those institutions that responded “Yes,” several of them specialize in acquisition of software.

Other more broadly focused collecting institutions stated that their agreement(s) included software only when it was part of the collection, but there was no standard text in their templates. For example one respondent noted, “When software is acquired it is reflected in our donor agreement, it is not part of a template.”

For most of the institutions (the other 82%) that do not currently account for software in their donor agreements or accessioning documentation, there was an even split between those that are planning to update their agreements, and those that are not.

Within the half that will update their agreements, 8.3% of respondents said they were planning to tackle this within the next year; the remainder, were more non-specific (“Yes, within the next few years”).

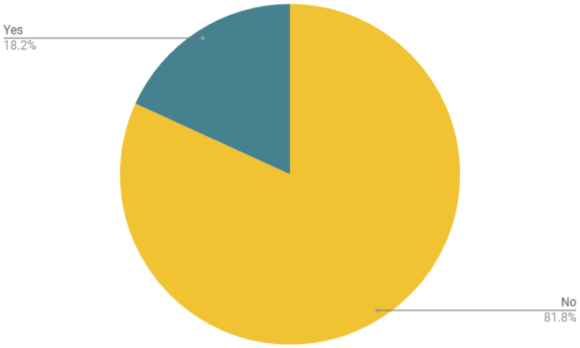


*Figure 4: Does your accessioning/donor agreement address software?*

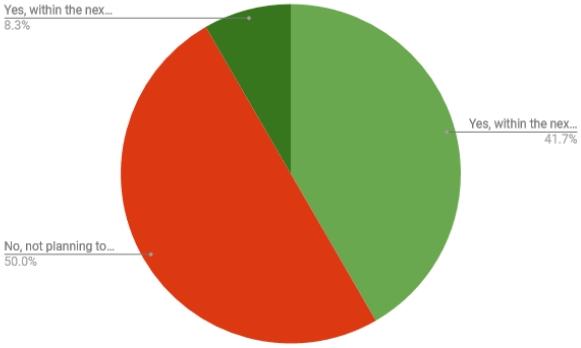


*Figure 5: Do you plan to update accessioning/donor agreement to account for software?*

There are some examples of donor surveys online that do include software:

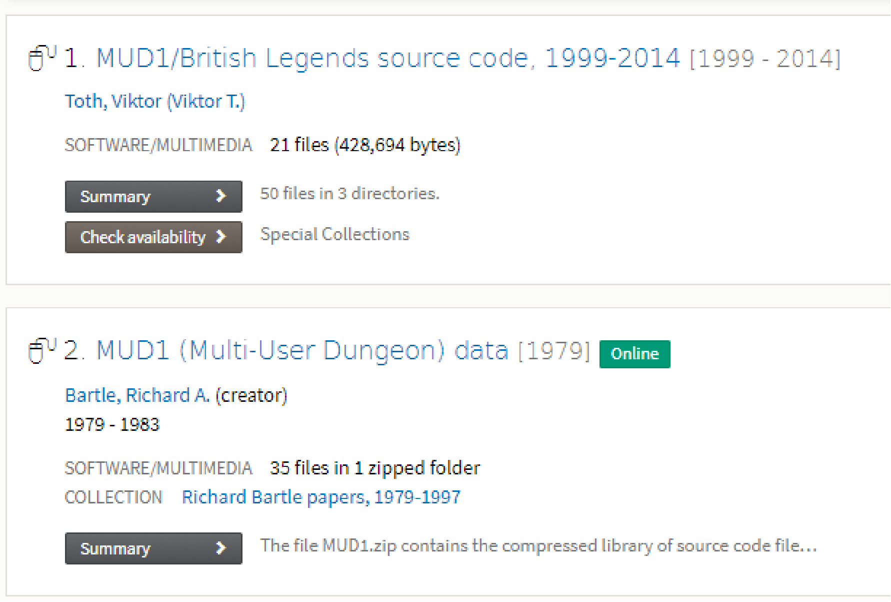
- Georgia Tech: [http://www.library.gatech.edu/archives/forms/GTSpecialCollectionsandArchives_DonorSurveyQuestions.pdf](http://www.library.gatech.edu/archives/forms/GTSpecialCollectionsandArchives_DonorSurveyQuestions.pdf)
- AIMS Project: [http://dcs.library.virginia.edu/files/2013/02/AIMS_final_appF.pdf](http://dcs.library.virginia.edu/files/2013/02/AIMS_final_appF.pdf)

However, we did not find any online documentation for donor agreements that included software.

Donor agreements that address software, whether or not it is part of their template, need to cover the same aspects as any other acquisition. They may restrict access or reproductions to reading room or research use only, or they may not. Nevertheless, this should be negotiated with the donor to ensure protection of their intellectual property.

For collections where the donor agreement documentation does not address software, the institution/repository might create another agreement or memorandum of understanding (MOU) regarding access and use of software. The following example is from the National Institute of Standards and Technology and Stanford University Libraries (NIST-SUL) project regarding copyright/access/capture of software titles in the Stephen Cabrinety collection. Recovering MUD1 software was part of this project.

There are two different versions of the MUD1 (Multi-User Dungeon) software at Stanford:


1. MUD1/British Legends source code, 1999-2014 [1999 - 2014]

Toth, Viktor (Viktor T.)

SOFTWARE/MULTIMEDIA 21 files (428,694 bytes)

50 files in 3 directories.

Special Collections

2. MUD1 (Multi-User Dungeon) data [1979] Online

Bartle, Richard A. (creator)

1979 - 1983

SOFTWARE/MULTIMEDIA 35 files in 1 zipped folder

COLLECTION Richard Bartle papers, 1979-1997

The file MUD1.zip contains the compressed library of source code file…

*Figure 6: NIST-SUL Multi-User Dungeon (MUD1)*

The MUD1/British Legends (Toth; 1999), which is still being maintained, is restricted to the reading room on campus; the second record, MUD1 (Bartle; 1979), is freely downloadable.

A good resource for additional information and other sample documents is the Association of Research Libraries' 2012 *SPEC Kit 329: Managing Born-Digital Special Collections and Archival Materials* (http://publications.arl.org/Managing-Born-Digital-Special-Collections-and-Archival-Materials-SPEC-Kit-329/). However, it should be noted that SPN's six-year roadmap includes activities to further identify, annotate and publish existing collection development, accessioning policies and donor agreement templates and guides for collecting software.

## *Do you preserve software?*

When asked if their institution preserves software, nearly 17 responded positively, while the majority (27) indicated they were not currently preserving software. Of those not currently preserving software, approximately half or 30% overall, indicated they would not begin preserving software in the foreseeable future. For those who are currently preserving software, their rationale varied from those who focused on software with enduring archival value to those that were preserving software platforms could render particular files. For those not currently preserving software, a common sentiment that one respondent articulated well is that, "We do not believe it is our purpose nor do we have the funds or resources to manage software, especially outdated software that will continue to pose security risks (with no current updates)." In nearly each case, it came down to the resources - both funding, staff time, and experience. Additionally, one repository stated that there was little administrative support for collecting and preserving software.

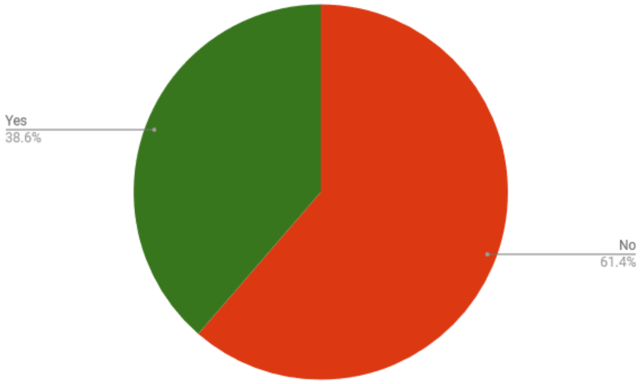


*Figure 7: Is your institution currently preserving software?*

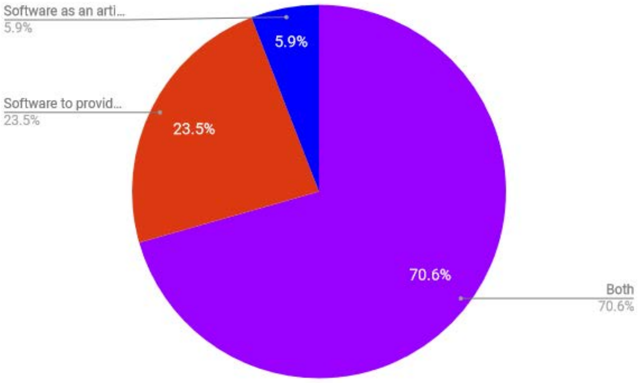


*Figure 8: Why are you preserving software?*

For those respondents who indicated that they are preserving software, one indicated that they were preserving it as an artifact, while four were preserving software solely to provide access to content. The other twelve, the vast majority, indicated it was for both purposes.

Perhaps an alternative strategy would be to create a universal registry or series of registries that collect executable software and maintain appropriate descriptive and technical metadata documentation.

## *Metadata currently used in preservation by institutions not collecting software*

For the institutions that are not currently collecting and preserving software, when asked, "What metadata is currently used in preservation?" the respondents identified thirteen distinct metadata standards, with many institutions using multiple types.

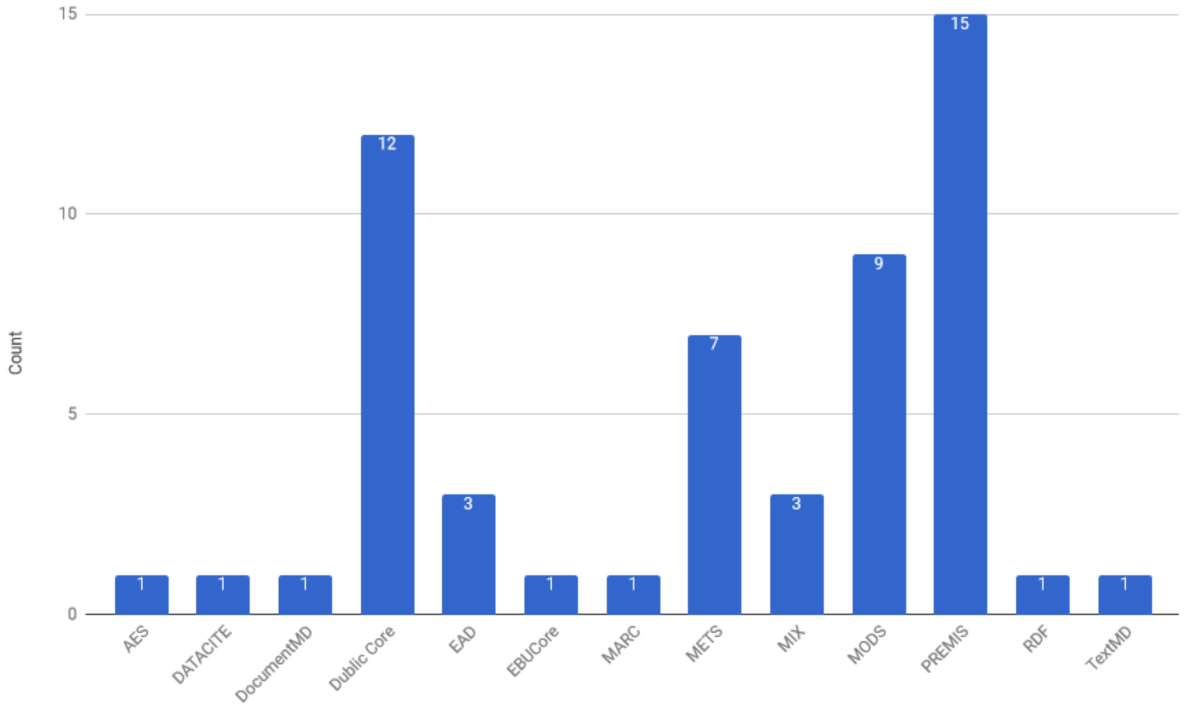


*Figure 9: Institutions not currently preserving software: What metadata are you currently using for digital preservation?*

In hindsight, the question might have been too broad as a one institution pointed out in the comments' "This feels like a very (!) broad questions, and so I'm not entirely sure that I know how to answer it accurately." Perhaps we should have separated technical, preservation, descriptive, rights metadata.

By far the most common were PREMIS (15) and Dublin Core (12) for preservation and descriptive metadata, with METS (9) and MODS (7) were not far behind. A few institutions mentioned in the comments which preservation software there were using, however this is tackled in more detail in a later survey question. The associated comments varied in perspective, and illustrate what many of us are grappling with. One institution listed the specific fields that they recorded including, file format, file version, file size, digital capture date and equipment, last modified date, and checksum. While another explained what went into their Archival Information Packages or AIPs - namely, PREMIS (preservation/technical metadata), DC (administration/rights md), BitCurator reports, photographs of legacy media, and relevant metadata gathered during the acquisition process, especially regarding any restrictions or redaction requirements. Yet another regional institution stated that they accepted whatever each participating institution choose - e.g. METS, MODS, etc.

A little over half of the respondents (24) answered the follow-up question, “Will this metadata be useful?” with nearly 61% indicating that the current metadata streams would be sufficient; while almost 9% said “no” outright, and approximately 30% undecided. One suggested that there is a need to develop new standards. The latter undecided group brought up some interesting points, namely, the need to record checksums to allow for comparisons against known software, and the need for tools to record or document dependencies required for preserving and accessing software.

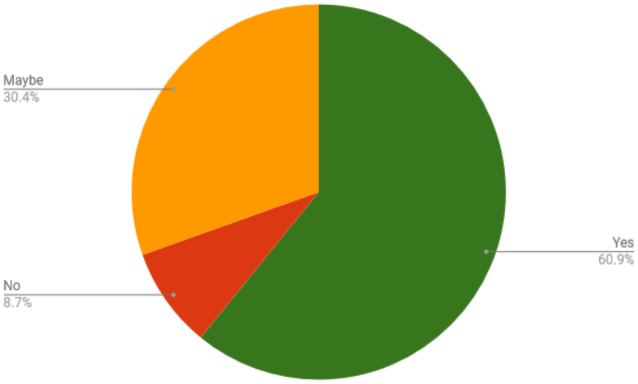


*Figure 10: Will this Metadata be useful?*

### *Metadata currently used in preservation by institutions collecting software*

For the institutions that are currently collecting and preserving software, when asked, “What metadata is currently used in preservation?” the respondents also identified thirteen distinct metadata standards (but not the same thirteen), with many institutions once again using multiple types.

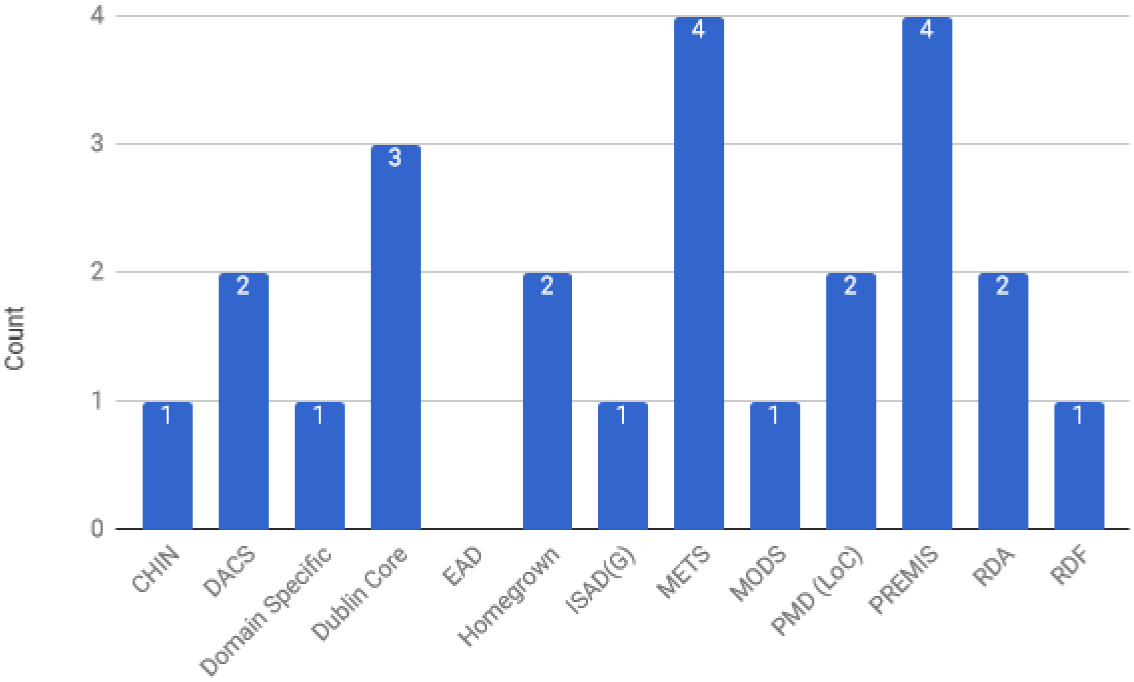


*Figure 11: Institutions currently preserving software: What metadata are you currently using for digital preservation?*

When asked what controlled vocabularies they were using in the preservation of software, 42% replied “None.” We might assume they are still in the planning stages. 26% stated they used home grown, domain specific or unidentified standards. Those few that did identify specific vocabularies listed, LCSH, RDA, Wikidata, and GAMECIP (media format and computer platforms). Interestingly, in a recently released NISO report discussed the complexity of the controlled vocabulary landscape. It identified the need for stability, documentation, and interoperability. [2] Without these elements, vocabularies are often orphaned.

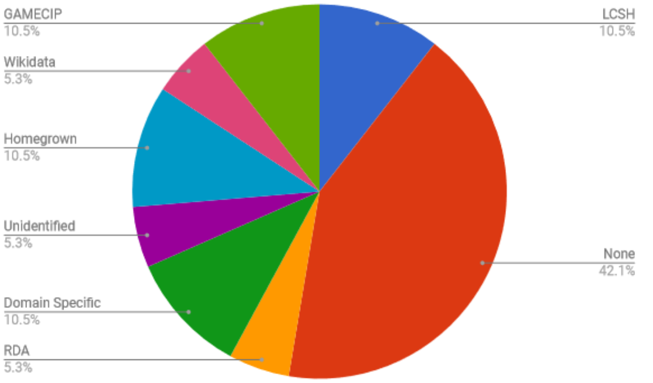


*Figure 12: Institutions currently preserving software: What controlled vocabularies are you using?*

Even more thought provoking, when asked what schemas or ontologies that they were utilizing more than 50% indicated “None”. The remaining responses were equally divided among domain specific, EAD, homegrown, METS, MODS, SWO, Wikidata and unidentified. Quite a few added in comments that they were currently recording minimal metadata.

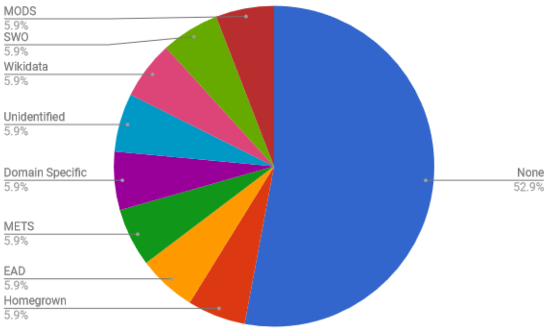


*Figure 14: What schemas/ontologies are you using?*

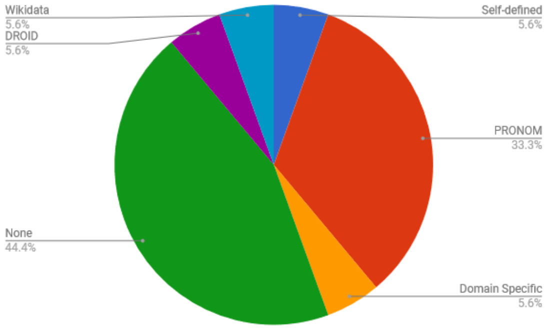


*Figure 13: What file format registries are you using?*

Further, when asked, “What file format registries are you using?” nearly half replied “None”, “n/a” or “?”. Of those that identified a format registry, the majority utilize PRONOM, while some indicated domain specific, self-defined, and Wikidata registries. One indicated DROID, which is actually a tool with links to PRONOM.

Finally when asked, “What systems do you use to preserve software?” the respondents listed a mix of more than twenty different preservation strategies, platforms and tools. The responses did highlight both where institutions were in the planning and process of preserving software as well as different strategies. Some institutions are incorporating Virtual Machines or emulation as preservation (and access) into their overall strategies; several gave a detailed list of tools, such as VirtualBox, DosBox, and WindowsXP emulators. While some only had local storage currently, others ingested software into a digital repository, some of which were internal/home grown, Archivematica, Fedora or Preservica.

*Table 1: What systems do you use to preserve software?*

| System/Tool | Count | Wikidata Item |
|---|---|---|
| Archivematica | 3 | https://www.wikidata.org/wiki/Q37610516 |
| BitCurator | 2 | https://www.wikidata.org/wiki/Q37761702 |
| bulk_extractor | 1 | https://www.wikidata.org/wiki/Q37760842 |
| CollectiveAccess | 1 | https://www.wikidata.org/wiki/Q2982932 |
| CVMFS | 1 | https://www.wikidata.org/wiki/Q28974795 |
| dd for imaging | 1 | https://www.wikidata.org/wiki/Q310581 |
| Device Side Data's FC5025 USB 5.25" Floppy Controller | 1 | https://www.wikidata.org/wiki/Q37762019 |
| DosBox | 1 | https://www.wikidata.org/wiki/Q479783 |
| FEDORA | 1 | https://www.wikidata.org/wiki/Q1400825 |
| fiwalk for analysis and documentation | 1 | none available |
| FTK Imager | 2 | https://www.wikidata.org/wiki/Q5468696 |
| Guymager | 1 | https://www.wikidata.org/wiki/Q37760071 |
| Institutional Digital Repository | 1 | none available |
| KVM | 1 | https://www.wikidata.org/wiki/Q377539 |
| Linux | 1 | https://www.wikidata.org/wiki/Q388 |
| None | 2 | none available |
| Preservica | 2 | https://www.wikidata.org/wiki/Q37621103 |
| RetroPi | 1 | none available |
| ScummVM | 1 | https://www.wikidata.org/wiki/Q145568 |
| Shelf storage for now/External Media | 3 | none available |
| SuperCard Pro | 1 | none available |
| USB-Floppydrive | 1 | none available |
| VirtualBox | 1 | https://www.wikidata.org/wiki/Q11393 |

## *Do you have additional metadata needs?*

The final question. “Do you have additional metadata needs?” was open ended and only ten institutions responded, which might be a result of fatigue, or the fact it was only answered by those that are currently collecting and preserving software. The prevailing thoughts are that:

- it is hard to agree on standards and solutions for metadata
- there are few if any good examples of metadata for software preservation
- there are also limits to existing standards and tools, i.e. not all the fields are included in each standard.

# Brief summary and future directions

The Working Group is very appreciative of the forty-four institutions that responded to the survey. From the responses, as seen above, we are able to draw a few general conclusions, and suggest many possible areas that need better documentation. First, most repositories seem to be in an early planning or contemplation phase–over 80% did not include software in the collection development policies or any specific language in their current donor agreements at this time. Furthermore, only a few repositories were planning to update their policies and agreements within the next year to account for software. Interestingly, even the institutions that are active collectors of software, with policies and agreements in place, are struggling with issues regarding metadata standards, best practice, and preservation workflows.

This survey reemphasized the need within the preservation community to continue to collaborate on standards, solutions, and tools for software preservation including:

- Centralized registries for:
    - software collecting and preservation institutions (archival value and/or for rendering files)
    - software collection development policies
    - agreements, language specific to software access, and rights documentation
    - obsolete software accessible to those that wish to render files
- Defining the limits to existing standards and tools—not all the fields are included in each standard—and developing a game plan for forward movement
- Defining a minimal set of metadata for software preservation, as well as outlining optimal sets that are situationally delineated (e.g. emulation)
- Documenting examples of existing metadata solutions for software preservation
- Considerations for including metadata as machine-actionable to support emulation as a service and other applications

The Software Preservation Network has already identified several of these areas for future investigation and development on its roadmap (http://www.softwarepreservationnetwork.org/category/news/). The Metadata Standards and Policies Working Group in addition to this survey and report are developing a crosswalk of existing metadata standards as they pertain to describing and preserving software. A preliminary draft of this effort is in GitHub.

# Appendices

## A. *Initial email requesting survey participants*

[Please excuse cross-posting]

The Software Preservation Network (SPN) is an initiative to explore and establish partnerships, collaborations, and best practices for software preservation. The SPN Metadata Working Group is currently conducting a survey of institutions with digital preservation programs to gain insight into metadata practices for software and other digital objects. The results of this survey will be used to establish a baseline for metadata best practices for software.

Please consider completing this very short survey about metadata practices in use at your institution. We are looking for responses both from institutions collecting software and those who have yet to begin this kind of work.

The survey is available at: http://bit.ly/SPN-MD-Survey until April 28**.**

You can follow our activities at http://www.softwarepreservationnetwork.org/working-groups/metadatastandards/

Thanks!
*The SPN Metadata Working Group*

## B. *List of questions*

Link to online survey: https://docs.google.com/forms/d/15V-fy8mKtLJDFXPD0_6hHgJYZXOjWvVH6OTDbyMF60s/edit

## C. *Electronic mailing lists original survey sent to:*

- SAA Electronic Records
- PREMIS Implementers
- Digital Curation Google Group
- SPN Mailing List
- Mashcat
- CLIR
- SAA Metadata & Digital Objects (MDOR)
- OLAC – ALA
- E-RECS Listserv (not affiliated with SAA)
- Digital Preservation Coalition
- Code4Lib
- Fedora Community
- Preserving and Archiving Google Group (PASIG)

## *D. Respondents' Job Titles*

Digital Preservation "category" covers:

- Digital services programmer
- Digital services librarian
- Digital preservation analyst
- Digital preservation librarian
- Digital preservation officer
- Data preservation project manager
- Digital services engineer
- System administrator
- IT service officer
- Service manager

Collection Services "category" covers:

- Digital archivist
- Digital processing archivist
- Digital collections archivist
- Electronic records archivist
- Electronic records specialist
- Head, special collections

## *E. Definitions*

- AES - Metadata standards for audio resources from the Audio Engineering Society.
- DataCite - DataCite is a non-profit organization working to promote data citation practices in scholarly communication.
- DocumentMD - Metadata schema for document formats.
- Dublin Core - The Dublin Core Schema is a set of vocabulary terms for describing web resources.
- EAD - Encoded Archival Description - XML standard for encoding archival finding aids.
- EBUCore - Metadata scheme containing descriptive and technical metadata for audiovisual resources.
- MARC - MAchine-Readable Cataloging standards are a set of digital formats for the description of items catalogued by libraries.
- METS - Metadata standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library.
- MIX - Technical metadata for digital still images.
- MODS - XML-based schema for bibliographic description.
- PREMIS - Metadata standard related to preservation of digital objects.

- PRONOM - online registry of technical information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.
- RDF - The Resource Description Framework is a family of specifications governing conceptual description or modeling of information for the web.
- TextMD - XML Schema that details technical metadata for text-based digital objects.
- Wikidata - Multi-lingual knowledge base of structured data.

## F. Resources

*Born Digital: Guidance for Donors, Dealers, and Archival Repositories Council on Library and Information Resources* by Gabriela Redwine, Megan Barnard, Kate Donovan, Erika Farr, Michael Forstrom, Will Hansen, Jeremy Leighton John, Nancy Kuhl, Seth Shaw, and Susan Thomas October 2013 https://www.clir.org/pubs/reports/pub159/pub159.pdf

*Collecting Software: A New Challenge for Archives & Museums* https://www.museumsandtheweb.com/biblio/collecting_software_new_challenge_archives_museums.html

---

[1] The information was self-supplied, not selected from a set of controlled vocabulary. There is room for overlap in our normalization process. For example, those that just identified as being part of a library may actually have been in a digital archives, digital preservation or archives and special collections unit.

[2] NISO TR-06-2017. Issues in Vocabulary Management: A Technical Report of the National Information Standards Organization