

Closing the Digital Curation Gap Focus Groups Report

Heather Bowden, Cal Lee, Helen Tibbo

Year Two Advisory Board Meeting

June 28, 2011 • London

Introduction

Many studies have been conducted to assess understanding and readiness of institutions charged with preserving and caring for digital collections. We have reviewed this previous research and have produced an annotated bibliography of over forty surveys of practice related to the management of digital information (Bowden, 2010). Previous research has done much to build a foundational understanding of the degree to which institutions understand and are prepared for the challenges that digital curation pose. We propose to build on this foundation by adding a greater understanding of digital curation activities as well as the tools, resources, and information sources that feed into these activities. Our goal is to use this information for the development of tools that will assist professionals in their digital curation needs.

We have designed and conducted focus groups with professionals working with digital collections as an extension of this previous research. In particular, it is an extension of results reported by the NEDCC (Clareson, T., 2006) and Cornell (Kenney, A.R. & Buckley, E. 2005). To build on this previous work, we assessed digital curation practices of professionals charged with the preservation and care of digital collections in library, archives, and museum settings. Particularly, we assessed what digital curation activities are currently being practiced, the tools and resources being used in these activities, where and how these tools and resources were discovered, and what kinds of tools and resources could help them better preserve and manage their collections. Additionally, we asked questions designed to determine if a decision tree tool would help professionals in their digital curation activities. The overarching goal of these focus groups was to get a sense of what practitioners are doing in the field and to learn where they need the most support and guidance.

Method

The focus groups have been designed to answer the following research questions:

1. What do the participants perceive to be digital curation activities?
2. In which digital curation activities are the participants currently engaged?
3. Which tools and resources are the participants using to aid in their digital curation activities?
4. Where and how did the participants acquire the tools and resources they use in digital curation activities?
5. How best would a digital curation decision tree tool assist in effectively practicing digital curation activities?

Participant recruitment was directed toward professionals charged with the preservation and care of digital collections in library, archives, and museum settings. For

the ALA and SAA focus groups, subjects were recruited based on their previous participation in digital preservation workshops and were contacted through listservs and lists of emails that were collected as part of these workshops.

We administered a total of four focus groups made up of between four and nine participants in each group, for a total of 25 participants. We held two focus groups in conjunction with the American Library Association (ALA) annual meeting. We had one group with six participants and one group with nine participants. We held one focus group in conjunction with the Society of American Archivists (SAA) annual meeting that was comprised of four participants. Our final focus group was held in conjunction with the Museum Computer Network (MCN) annual meeting and had six participants.

The focus groups were led by Cal Lee, while Heather Bowden and Helen Tibbo took notes. The focus groups were also recorded using a digital audio recorder during the sessions.

The focus group discussions were guided by the following questions:

1. What kinds of activities do you believe are involved with digital curation?
2. What kinds of digital curation activities do you currently practice?
3. For these activities, what tools and resources have you used?
4. Where and how did you find these tools and resources?
5. What other types of tools and resources would you find to be helpful?
6. How might you use a decision tree (such as this)?
7. What would you see as the main opportunities or benefits of using such a decision tree?
8. What would you see as the main challenges in trying to make use of such a decision tree?
9. What else should we have asked you?

The chart shown in Figure 1 was printed and handed out to the participants before the focus group sessions began and was used as a question guide throughout the sessions.

	1. Planning	2. Create or Receive Content	3. Appraise, Select & Reappraise Content	4. Ingest Content	5. Creation and Management of Metadata	6. Take actions to preserve content (e.g., migration, emulation)	7. Store Content	8. Manage Access, Use and Reuse of Content	9. Disposal of Content
A. Curation Activity Currently Practiced	1A	2A	3A	4A	5A	6A	7A	8A	9A
B. Tools & Resources	1B	2B	3B	4B	5B	6B	7B	8B	9B
C. Source of Tools & Resources	1C	2C	3C	4C	5C	6C	7C	8C	9C
D. Desired Activities, Tools & Resources	1D	2D	3D	4D	5D	6D	7D	8D	9D

Figure 1. Question matrix handout used to guide the focus group discussion. See results below for highlights of the interviews based on the matrix.

At the conclusion of the discussion centered on the question matrix, we displayed to the participants, either on a projected screen or on a color printed handout, two views of a conceptualization of a digital curation decision tree. These conceptualizations are shown in *Figure 2* and *Figure 3*. The discussion guided by questions 8 and 9 were centered on these proposed decision tree views.

Digital Curation Resource Tree

> Plan > Have you created a comprehensive digital curation policy? > No

Choose a digital curation activity

Plan

Create or Receive Content

Appraise, Select,
and Reappraise Content

Ingest Content

Creation and Management
of Metadata

Take Actions to Preserve Content

Store Content

Manage Access, Use,
and Reuse of Content

Dispose of Content

Create a Digital Curation Policy

Read.

- Au Yeung, Tim. Digital Preservation for Museums: Recommendations. Commissioned by the Canadian Heritage Information Network. Gatineau, Quebec, Canada: Minister of Public Works and Government Services, 2004. [\[PDF\]](#)
- ERPANET Digital Preservation Policy Planning Tool [\[PDF\]](#)
- Library of Congress. National Digital Information Infrastructure and Information Preservation Program. Sustainability for Digital Formats: Planning for Library of Congress Collections. [\[Link\]](#)

Review Examples.

- Columbia University Libraries. Policy for Preservation of Digital Resources, July 2000, revised 2006. [\[Link\]](#)
- Cornell University Library Digital Preservation Policy Framework [\[Link\]](#)
- ICPSR Digital Preservation Policy Framework Model [\[Link\]](#)
- National Library of Australia [\[Link\]](#)
- University of Illinois at Champaign-Urbana. IDEALS Digital Preservation Policy [\[Link\]](#)
- Yale University Library [\[PDF\]](#)

Take Action.

Get instruction.

- Attend DigCCurr Institute [\[Link\]](#)
- Attend the Digital Preservation Management Workshop [\[Link\]](#)
- Review materials from the Joint DCC/DPC Workshop: Policies for Long-Term Curation and Preservation [\[Link\]](#)

Create your digital curation policies.

Figure 2. Conceptualization of a view on creation a digital curation policy for a proposed digital curation decision tree.

Digital Curation Resource Tree

> Take Actions to Preserve Content > Manage File Formats > Can you choose which file formats you ingest? >Yes

Choose a digital curation activity

Plan

Create or Receive Content

Appraise, Select, and Reappraise Content

Ingest Content

Creation and Management of Metadata

Take Actions to Preserve Content

Store Content

Manage Access, Use, and Reuse of Content

Dispose of Content

Choose Preservation File Formats

Read.

- Judith Rog and Caroline van Wijk. Evaluating File Formats for Long-term Preservation [\[PDF\]](#)
- The National Archives. Selecting File Formats for Long-term Preservation [\[PDF\]](#)
- File-Formats for Preservation: Evaluating the Long-Term Stability of File-Formats [\[PDF\]](#)
- Library of Congress. Formats, Evaluation Factors, and Relationships. [\[Link\]](#)

Review Examples.

- California Digital Library [\[Link\]](#)
- Harvard University Library [\[Link\]](#)

Take Action.

Get instruction.

- Attend DigCCurr Institute [\[Link\]](#)
- Attend the Digital Preservation Management Workshop [\[Link\]](#)

Find tools.

- Planets Testbed [\[Link\]](#)
- Plato [\[Link\]](#)

Choose the file formats to be ingested into your collection.

Figure 3. Conceptualization of a view on creation a digital curation policy for a proposed digital curation decision tree.

Results

It is important to note that though we recruited participants in conjunction with ALA, SAA, and MCN annual meetings, the individual focus groups were not composed homogenously of librarians, archivists, and museum practitioners. Each focus group was comprised of members of more than one “category” of practitioner. This is not altogether

surprising as it shows that while there are very distinct differences between the three fields, there is still a certain degree of overlap between them.

The process of conducting the focus group interviews also pushed categorical boundaries. While the question grid was used to direct the interview process, the discussions were far from linear and categorize-able. Much of the discussion was general and did not adhere strictly to the categories that we used to guide the conversations. Consequently, many of the answers that were recorded do not fit under a particular activity category, but rather are more general statements about the participants' thoughts about their digital curation work. We did take notes of the conversation based on the flow from category to category and have collected quotes within these categories where we could.

Additionally, we realized very quickly that the sources the participants consulted for guidance and tools were relatively the same for each activity. This is reflected in the slow tapering off of quotes for the questions related to the sources of tools and resources. Google, listservs, conferences, and other practitioners were the most prevalent sources of information, and it was indicated throughout all of the focus groups that these sources were consulted for each of the digital curation activities discussed. The discussion of the tools and resources and their sources also naturally blended together and are combined in the highlights of quotes from the notes below.

Highlights from the notes below follow the numbering system in *Figure 1*:

1A. Planning Activities Practiced. "Not many people practicing formal planning." "Generally ad hoc." "We are not planning at all." "It's all project oriented and even those aren't planned out." "They are just scanning things and then adding metadata." "Project planning is a struggle for us to say the least." "We are in a planning process, but I feel like we are going to be there forever." "Our plan is to come up with a better plan."

1B & 1C. Planning Tools & Resources. "Online, listservs, conferences" "Agile Scrum" "Other institutions and what they have published." "Other people." "DCC blog, Fedora learning community blog." "ContentDM has provided for all of the libraries that are members." "CLIR reports, Dspace and Fedora community." "DLF "Digital Library Federation." "Meetings, workshops, training." "Aside from colleagues anything that's online." "UNC, UIUC, and now Arizona have people who know things."

1D. What kinds of planning activities would you like to practice in the future and what other kinds of tools and resources would be helpful? "An anonymous database of failures where we can learn what people have done and failed." "An outline of milestones that we would be shooting for and then to see where people have fallen short on those milestones" "More funding" "Education for managers and elected officials. They think if they have back up they have digital preservation" "Digital disaster recovery, specific to our region."

2A. Creating or Receiving Content Activities Practiced. What do you do? "Digitization off-site." "Internet Archive to capture websites." "Automate transfer of electronic records from state and local agencies." "Working with born digital records, but getting a lot of digitized." "Scanning microfilm." "Developing workflow for structured database records."

"We base some of what we are doing on the paper model even though we are not making documents on paper so much." "We don't get our documents right after they are created. We get them like 25 years after they are created and it can be imperfect." "Getting CIP data from LC and ONIX records"

2B & C. Creating and Receiving Content Tools & Resources. "Archivists toolkit and Bagit, to implement transfers" "Microsoft Biztalk, for process control." "Studies by AAM and Cornell that deal with copyright. Also a lot of listservs in this community and institutions that have tackled these problems." "For art museums: College Art Association, Visual Resources Association, Art Librarians of North America" "AIIMS." "Minnesota's trustworthy information Handbook"

2D. Creating and receiving content activities you would like to practice in the future and what other kinds of tools and resources would be helpful? "Some easy way for institutions to manage and share their resources. It's really hard if it's not uniform." "A big hurdle is researching the copyright status of images before you put them online." "A lot of institutions don't have in house council for dealing with legal issues" "Supported tools and functions for preservation." "Something that can do everything. I don't care if it's some monolithic structure." "Better awareness campaign about what the library can do."

3A. Appraisal, Selection, and ReAppraisal Activities Practiced.

"We have a form that we modeled after Harvard; that before someone submits something, we have them fill it out." "It's currently as needs come up. It's based on events and activities." "Standard business classification scheme to help with appraisal and selection of content up front." "Macro appraisal." "Domain selection tool for ArchiveIt." "Make collection and appraisal decisions based on space."

3B. Appraisal Tools & Resources. "We looked to Harvard, also North Carolina." "Fedora." "What are other industries doing? It's not just archives, but also business. What about people who have more than just business documents. "

4A. Ingest Activities Practiced. "Putting things in DSpace is the closest." "Bagit, to make sure it's not corrupted." "One of our challenges is bandwidth because we're dealing with video." "I still treat it like any other accession. List where it is, "shelf location", or on spinning disc. It's still ad hoc." "We run checksums, but that's it." "We rely on Fedora, and checksums, if they provide a checksum, Baggit specifications"

4D. Ingest activities you would like to practice in the future and what other kinds of tools and resources would be helpful? "A tool to automatically connect provenance. When an office changes names, that is tracked. Visualization of the links different offices are associated with and be able to easily attach" "We are spreadsheet oriented, and we need to get away from that. Ingest is pretty clunky." "Hit 'scan this book' and you hit the bar code and it automatically pulls the metadata and does the format."

5A. Creating and Managing Metadata Activities Practiced. "Metadata is the critical thing. It's the most expensive thing. We can fund digitization, but it's almost impossible to

catalog from scratch.” “MicroMARC. Is what we are using now, but migrating into Archivists Toolkit.” “Everything we’ve done is in MARC” “We all use a metadata document that we’ve posted to the wiki.” “Metadata isn’t a term used in the museum world.”

5B & C. Creating and Managing Metadata Tools & Resources “PB core. We’ve identified a whole lot of problems with it when moved into a digital environment. Move to PB core 1.0 to 1.5 which was acceptable enough. We had to lobby for PB core 2.0, which would be acceptable” “Forced to go into Dublin Core.” “We rely on Dublin Core. We do item level description and it’s the first thing we do before we scan.”

5D. Creating and managing metadata activities you would like to practice in the future and what other kinds of tools and resources would be helpful? “We need something to help us... don’t know what that is.” “Anytime you bring up anything about computers, I think a lot of my colleagues feel lost and don’t even want to approach it.” “List of resources, workshops at that basic level that would be really helpful.” “We don’t have one tool that’s telling us that people are searching for things in ways that are using the metadata.”

6A. Preservation Activities Practiced. “Committed to text only platforms so I don’t have to worry.” “It’s very much ad hoc.” “What we’re doing is not worrying. It’s in formats that we expect to be around for 5-10 years.” “All we’ve done is back up and fixity and I don’t know that we’ve even done fixity.” “We haven’t had to go through any migration yet. We can identify every file that is of a particular file type made at a particular time period and we have gone through the process of transcoding them for output to a different kind of device, but no real migration.” “Trying to see what people are doing and how they implement it, but there are a not a lot of people who are doing it.” “We haven’t even tried to wrap our heads around it.” “We’ve started the digital curation planning team.” “It was supposed to be a preservation environment, and then it turned into an access environment. People just want to see a pretty website and we are losing the underlying preservation goals. We’re relying on IT staff to maintain backup mechanisms.” “Microservices at CDL (California Digital Library).”

6B & C. Preservation Tools & Resources. “I think this is something that vendors will have to do for us, because as we scope this out I think this is outside of our purview.” “I think it comes back to what are the tools to do this?” “FRED machines. Forensic Recovery.” “Computer Preservation Society – working on preserving programs, formats, and operating systems.” “I’m excited about Interpares 3.”

6D. Preservation activities you would like to practice in the future and what other kinds of tools and resources would be helpful? “We need easier tools. I’m not a programmer and I’m trying to use JHOVE and it’s difficult. Having an easier tool to help with preservation metadata.” “Registry of different formats, but maintained.” “There are motivated enthusiasts to move these things forward. We can work to connect with these groups to move things forward.” “Have some kind of workshop where you’re there for at least a day and have higher level and lower level person from the same place talk these things out. Have a case-study of a place that has worked these things out.” “I would like to

see a map that lays it all out." "I envy the Europeans where there are mandates to do these things."

7A. Storing Activities Practiced. "Courier service to move tapes around as "distributed, offline backup/storage" "Looking at cloud storage within consortium - Also use DSPACE on university server" "Developing tiered storage with different levels of curation." "16 TB LOCKSS boxes." "Renting/purchasing space on server." "Live internal network and backing up on gold CDs, which are a nightmare." "We're just putting things on DVDs, but they're all over the place." "We haven't put anything on an offsite database. People don't feel confident enough to do it." "Storage is such a huge issue. We're being held hostage by the systems department and we don't speak their language, we just tell them we need space."

7D. Storing activities you would like to practice in the future and what other kinds of tools and resources would be helpful? "We keep hearing that storage is cheap but we hear back from IT that it isn't. We understand the cost of storing an artwork, but it would be interesting to know the cost of preserving digital content." "If I had resources to organize everything in a central repository instead of everyone having to be on different hard drives, I think we could save a lot of money in storage. It would be an ongoing process of building trust." "Dealing with barriers to working with IT." "We need to make a case for why these things are a priority."

8A. Access Activities Practiced. "We use ContentDM. It's incredibly expensive. It's a hosted service and our sys admin is not very proficient. We are moving to Drupal hopefully soon." "We have been using Flickr. It's been far more useful than ContentDM. It's generated far more hits. But content creators don't want to see their content in Flickr, because of control." "One thing that's been really challenging having people access things is bandwidth" "Working with KORA - open source software, to provide access to digital surrogates. Might look at Omeka - Luna, image management system"

9A. Disposing Activities Practiced. "It depends on how you define the collection. There are a lot of photos that are not part of the collection, that are absolute garbage, but people save them. I'd love to get rid of some things." "We scanned someone's prints and they were tossing the tiffs and keeping jpg derivatives because they didn't want to pay for storage." "Shred the drives" "Possibly use Archivists toolkit" "We have to keep our records for the life of the site plus six years." "This one applies to us in that we were looking at old digital collections to see if they were salvageable. We had to decide if we wanted to recreate it or work with an old collection." "If you migrate a file we keep the original, so we do not dispose of the first ingest."

Decision Tree comments. "In addition to the column to have a space for sharing experiences and examples of this work in action." "Museums like to see museums represented." "Because things change so often it would have to be more dynamic and be changed by the user community." "Institutions are interested in what similar sized institutions are doing, so tier information for size." "This is also begging for marginalia for people to chime in. Comments off to the side." "It's a good tool for research. My biggest problem is getting people above me to understand what needs to be done." "The more

specific it gets toward the end, the better it will be. When you are trying to educate yourself and you get to the very specific thing is the hardest.” “If you could have it as user ranked. I would have a better idea of what other people are using and like.”

Discussion

One of the most useful discoveries we made in the process of conducting these focus groups is the diversity of scenarios in which digital curation activities are undertaken. Digital collections are being managed in very small institutions as well as larger ones. It is being taken on by those who have little to no experience with technology and by those who are trained computer scientists. This range of practice experiences was mirrored by the range of participants who took part in the focus groups. We had participants who had little to no experience or deeper understanding of the full digital curation lifecycle and we had participants who were well versed. While we focused on practitioners coming from small-to medium-sized institutions, we had participants from a diverse range within this subsection.

It is worth noting, the differences and likeness we discovered between and among those participants from museums, archives, and libraries. Much of the lexicon that is used in digital curation stems from that of archives and is often more directly applicable to the experiences of archivists. It was noted in one of the focus groups that the practices of libraries and archives are more similar and many of the terms cross over from archives to libraries without much confusion. One participant from a museum setting noted that they do not use the term metadata regularly in museums. Additionally, they do not refer to the materials in their collection as “content.” These differences could have implications in the construction of decision trees and other resources if they are to be made useful across all three practicing fields. This should be explored in more depth as we progress with the decision tree development.

Throughout all of the focus groups, it was surprising that the participants reported that their institutions had little involvement in formal digital curation planning. Many indicated that their institutions wanted to do more formal planning, but that most of their digital curation activities were performed ad hoc. One of the most notable quotes from the planning discussions that exemplifies this is, “Our plan is to come up with a better plan.” Not surprisingly, the discussion content shows the strongest foundation in the creation and acquisition of content and providing access to their collections.

Almost all of the focus group discussions indicated that while storage of their content was very important, building and maintaining relationships with the IT staff that manage storage can be very challenging. Additionally, almost everyone was able to talk about what they do for metadata, but the creation and management of their collections’ metadata is very expensive. Preservation activities are still an area of great concern. When we asked one focus group about their preservation activities, the group responded by laughing and shaking their heads. The biggest concern for the participants is that there are not yet any tried and tested tools that are available. One participant indicated that the tools that are available are not very user friendly and require a certain level of programming skills.

All of the focus group discussions about the decision trees were very useful. Many of the participants indicated that these should not be static web pages, but built in a dynamic

system that could be continually updated as things changed. They also indicated that they would like the ability to leave comments or even rate certain resources that were listed in the decision trees. Several focus group participants indicated that having institution specific information would also be helpful, for example resources specific to institution size or type, i.e. museums, libraries, or archives.

Conclusion

The primary goal of these focus groups was to gain a clear understanding of what digital curation practitioners are doing successfully and where they need guidance and support. Most importantly, we want to build this understanding into a resource that practitioners can use to guide them through the many processes of digital curation practice. We have started with a conceptualization of what we called, “Digital Curation Decision Trees” and based on the information collected in these focus groups we have set out to design a tool that can address these needs. Because there was a consistent call for interactivity and evolve-ability in such a tool, we decided to begin building it in the open source content management system of Drupal and have it linked to the resources that we have already started collecting in the Drupal-based Digital Curation Exchange website.

We have also given the name “Decision Tree” some consideration. As we have begun building the structure of this resource, we have realized that the information architecture held little equivalency with true decision-tree logic. We are considering new names for this resource, such as a “Getting Started Guide (GSG),” or a “Get Going Guide (GGG).” Further development of the “Guide” has been informed by one-on-one interviews with practitioners. Each of these interviews is focused on one digital curation activity such as archiving websites, building an institutional repository, digitizing collections, managing and preserving scientific data, and pulling digital information from physical media. The information we gathered from these interviews is being used to shape individual guides for each of these activities. Further interviews will be held through the life of the project and will be used to inform continued development of the Guides.

References

Bowden, H. (2010). Closing the Digital Curation Gap Annotated Bibliography of Digital Curation Practices. Retrieve June 15, 2011 from http://ils.unc.edu/gapCDCG_AnnotatedBib.pdf.

Clareson, Tom. (2006). NEDCC Survey and Colloquium Explore Digitization and Digital Preservation Policies and Practices. *RLG DigiNews* 10(1).

Kenney, A.R., and Buckley, E. (2005). Developing Digital Preservation Programs: The Cornell Survey of Institutional Readiness, 2003-2005. *RLG DigiNews*, 9(4).