NDSA Standards & Practices Interest Group

2024 Rolling Agenda and Meeting Notes

Co-Chairs: Michael Dulock (michael.dulock@colorado.edu), Ann Hanlon (hanlon@uwm.edu)

Join Zoom Meeting

https://clirdlf.zoom.us/j/87021224128?pwd=G19JbJ6WmpoqivK44BoXxYgCynJ9gb.1

Meeting ID: 870 2122 4128
Password: DLFzoom
One tap mobile
One tap mobile:
+13092053325,,87021224128# US
+13126266799,,87021224128# US (Chicago)

Dial by your location.
Find your local number: https://clirdlf.zoom.us/u/abZgulOJO

Meeting Dates

- 2024: Tuesday, January 9th; Monday, April 1; Monday, July 1; Monday, October 7
- 1 pm Eastern

Old Meeting Notes

- 2023 Meeting notes
- 2017-2021 are available in the Google Drive Folder

Add to calendar:

- Please download and import the following iCalendar (.ics) files to your calendar system.Monthly: https://clirdlf.zoom.us/meeting/tZMtdeiqrD8rGNyJzstYFqK2Pdp4oiARJucB/ics?icsToken=98tyKuGvrjkrGNaRsRuPRpwEAojod-rzpilcjY1EtgX2Fxl1cyikBuZSZr12Mer6

NDSA Standards & Practices folder:

https://drive.google.com/drive/folders/1c3D1RI1DQGmoKI5Y-9CzSBjUBq9Z3sYK?ths=true

NDSA Slack

- Workspace Link
- Join Link
    - You can add yourself to any of the public channels including one for the Standards and Practices Interest Group.
    - Use this to communicate and collaborate with others within NDSA.
    - NDSA Slack User Guide

NDSA Code of Conduct

NDSA groups follow Digital Library Federation's (DLF) Code of Conduct as the Council on Library and Information Resources (CLIR) acts as the host organization for both DLF and NDSA. Website with full details: https://ndsa.org/about/code-of-conduct/

If an incident occurs during an Interest Group meeting please use the method below that is most comfortable for you.

- Reach out to the identified co-chair or appointed code of conduct monitor via private Zoom chat during the meeting.
- Reach out to any of the Interest Group co-chairs after the meeting.
- Report the Code of Conduct concern/violations using the anonymous form. This form is received by the Chair and Vice Chair of the Coordinating Committee.
- Report the Code of Conduct concern/violations to conduct@ndsa.org. This email is monitored by the Chair and Vice Chair of the Coordinating Committee.
- Report the Code of Conduct concern/violations to one or more people on the Leadership team

---

2024-10-07

Agenda: Persistent Identifiers and Preservation

Persistent identifiers, or PIDs, as the name suggests, are intended to provide long-term access to a digital object. Given the importance of access and use to the preservation of digital objects, S&P proposes a discussion about how and when we use PIDs in our digital preservation activities – if we are using them.

Do you use them at all? And if you do, what policies and practices guide your implementation? Join us for our quarterly meeting on October 7th for a discussion about PIDs.

Present: Ann Hanlon, Michael Dulock, Michael Barera, Ima Oduok, Michael Barera, David Larsen, Shana Scott, John Kunze, Dina Sokolova, Hannah Tashjian

Discussion notes: Ann and Michael have thought about PIDs in context of migrations. How to maintain access to objects as they move to different platforms.
John Kunze in attendance gave us a 10-minute presentation about PIDs (slides and notes). Myths about persistent identifiers. n2t.net/ark:/13030/c7gb.
PIDs are permalinks that may be recognizable. “All PIDs are aspirational.” One important tool is redirection. PIDs break often, and content providers repair them. Direct, redirect.
PIDs and resolvers - all similar. Resolver | Name assigning authority | name.
Why use PIDs at all? The recipient knows it has a higher chance of resolving and may be able to query for metadata services, etc. As it happens, each ARK has a globally unique identity that doesn’t depend on domain name.
Boulder is doing collection and item level ARKs; cultural heritage, archival, and rare books, as well as geolibrary are the content they use ARKs for.
Kunze suggests it doesn’t have to be a binary decision which identifiers to use - normal to use more than one scheme at once. Some folks assign ARKs to all as it's free, and assign DOIs to published stuff since that’s customary, etc.
Boulder has documentation and workflows, but no policy.
David Larsen said they have never used URIs, but are looking to do something now. The Africa PID Alliance has been kicked off in his area and looking to connect to that. There is a similar effort in West Africa, as well (WACREN).
Question of policy? What kind of commitment are we (the repository) making to this thing? Reasonable commitment versus long-term commitment. What do we mean by persistence? One kind of persistence is to a site that changes but is the same (NLM example) versus “unchanging” bitstreams (docs, legal docs, etC); we're replacing bitstreams all the time with something potentially better. Agreeing on what is “persistence”.
The Global South has been using ARKS for publications, so it’s now becoming known as an identifier for publications.
Migration will always mean “shock” to the user in terms of the URL they are using (ie, changing from old ids to new ids). There will always be a table to manage with redirects.

2024-07-01

Agenda: A discussion about selection for preservation

What practices and policies do we have in place to guide selection for preservation? And how are we taking environmental impact into account? What factors matter most when we are prioritizing which materials receive the most attention. Join us to discuss and share practices across our institutions. And of course, feel free to refer to the NDSA Curatorial Decision Guide: https://osf.io/fz98s

Present (name and institution)

- Ann Hanlon, UW-Milwaukee, Margaret Turman Kidd, VCU, Carol Kussmann (UMN), Allison Look (Illinois Institute of Technology), Peter Gorman (UW–Madison), Shana Scott (Anderson Archival), Alice Prael (Yale Libraries), Kimberly Kennedy (Northeastern University), Dina Sokolova (Columbia University), John Kunze (Drexel University|ARK Alliance)

Agenda

- Update on Cloud Services and Digital Preservation group
    - Subgroup formed within this group. Met with people from the Infrastructure group/storage survey group. Waiting to get de-identified data from the storage survey group - should be soonish.
- Selection for preservation discussion. What policies guide digital preservation?
    - Digital curation decision guide Levels of Digital Preservation
    - UW-Madison - not everything needs the highest level of digital preservation. Adapted Penn State's guide for assigning level (research value, scarcity, size, formats…) gave a rubric/number to determine what level of preservation is needed. Highest gets added to the preservation system. Fedora repository is level 0, preservation archive is different. (Wi version)
    - UMN - Preserving the materials that we create. Materials from the Minnesota Digital Library. Data Repository is at the 10 year mark, thinking about what a criteria to keep these but no plans to move forward with removal at this time.
    - UW - Milwaukee - Born digital and legal and unique are the things that are being targeted for immediate preservation. Policy. Need to appraise as space is limited.
    - How does environmental sustainability/factors and/or appraisal come into play with these decisions?
        - Trying to balance environmental sustainability and the recommended number of copies for preservation. Ex. Working with APTrust who supports the replication, so probably don't also need a copy locally.
    - How to make time to go through larger collections to understand what the content is during appraisal. Have some archivists (more than before) doing appraisal. Work is becoming more collaborative. Have a born digital community of practice to support work in general (weekly co-working hour). Involve more people - subject experts as well as the people on the technical side.
    - Don't always do a lot of re-appraisal, making the process more difficult. (looking at content on media, time consuming)
    - Appraisal as part of the accessioning/processing process involved thinking about what we actually need to preserve and at what quality level.

    - ARK (Archival Resource Key) - concept of persistence statement (what do they mean by this term). This can be different for different orgs, collections, objects.
        - https://doi.org/10.5334/dsj-2017-039 Persistence Statements: Describing Digital Stickiness (2017) -- In this paper we present a draft vocabulary for making "persistence statements." These are simple tools for pragmatically addressing the concern that anyone feels upon experiencing a broken web link. Scholars increasingly use scientific and cultural assets in digital form, but choosing which among many objects to cite for the long term can be difficult. There are few well-defined terms to describe the various kinds and qualities of persistence that object repositories and identifier resolvers do or don't provide. Given an object's identifier, one should be able to query a provider to retrieve human- and machine-readable information to help judge the level of service to expect and help gauge whether the identifier is durable enough, as a sort of long-term bet, to include in a citation. The vocabulary should enable providers to articulate persistence policies and set user expectations.
        - Group looking at creating the minimum viable product - the 80%. Defining what we will or won't do. Helps with user expectations as well.
- Topics to explore next

2024-04-01

Agenda: The Language of the Cloud

Let's help each other understand the terms that we grapple with when we outsource infrastructure to major cloud-based technology vendors like AWS and Azure. Come prepared with terms and concepts that are distinctly cloud-based and that impact your practice, and we'll discuss what they mean and why they are important to understand.

Present (name and institution)

- Ann Hanlon, University of Wisconsin-Milwaukee
- Michael Dulock, University of Colorado Boulder
- Margaret Turman Kidd, Virginia Commonwealth University
- Hilary Wang, Brown University
- Kim Gianfrancesco, Vassar College
- Patrick Daglaris, Oklahoma State University
- Tyler Thorsted, Brigham Young University
- Elizabeth Paris, Anderson Archival
- Allison Look, Illinois Institute of Technology
- Ima Oduok, Texas Digital Library
- Lindsey Memory, Brigham Young University

- Dan Noonan, The Ohio State University
- Ling Meng, University of Wisconsin-Milwaukee
- Scott Prater, University of Wisconsin-Madison
- Peter Gorman, University of Wisconsin–Madison
- Stephen Abrams, Harvard University
- Bethany Scott, University of Houston
- Krista Oldham, Texas A&M University
- Fatemeh Rezaei, The University of Baltimore

Pondering Standards & Practices related to how we interact with The Cloud. Standard language, working with “brokers” (Preservica, APTrust) or direct vendors (AWS, Azure).

Variety of backgrounds in Digital Preservation work. What are common questions/understandings/misunderstandings?
Can be opaque - talking to broker/vendor through institution IT liaison? Directly?
Coming up to speed on service structure, service offerings & differences, charges related to various actions/activities, how to calculate, etc.
Can be hard to know what questions to ask.
Harvard: Enterprise contract w/ AWS, single-purpose contract through Libraries w/ Wasabi (looking at enterprise contract w/ sub-contracts for various units/use cases).
Internally, storage calculations based on true TB. But AWS, for instance, works in tebibyte (binary- rather than decimal-based, 10% bigger).

> https://aws.amazon.com/s3/pricing/

Non-commercial cloud providers like MetaArchive?
Ann & UW-Milwaukee using Preservica
APTrust (CU Boulder, VCU) some info: https://aptrust.org/about/storage-fact-sheet/
Harvard keeping min. 1 copy of everything on their own storage infrastructure. CUB is unofficially but in practice doing the same thing (no policy in place). UW-Madison & UW-Milwaukee, Vassar, & Northeastern doing same. OSU moving to cloud (AWS enterprise contract), dropping local storage for cost (backups!). BYU local tape & one copy in BYU-controlled cloud.
Harvard & local Research Computing, inexpensive - some difficulties since our use cases are different. Some reconfig work was needed to ensure isolation between clients. CUB using similar local service, not the typical use case.
Digital Preservation as a core principle of Libraries/higher ed institutions. Is local and/or cloud storage recognized as a core need for library preservation?
What about smaller vendors like pcloud or idrive?

> Pcloud in use some for personal use, there is a “lifetime” purchase available. There are business solutions but not directly investigated.

DLF blog post in this neighborhood:
https://www.diglib.org/using-cloud-storage-for-access-to-digital-archives/
What are our practices & are they changing?

For next meeting, if this is a good avenue to pursue: Pull out main points, look for a guest to discuss?

2024-01-09

AGENDA:

- ***Digital Preservation system migration***: Considerations and Lessons Learned: Michael Dulock (University of Colorado Boulder) will discuss CUB's migration from the Arkivum digital preservation solution to APTrust, including migrating indirectly, pull-down/push-up style, and the intermediate systems used to accomplish that. He will talk about what went into CUB's decision to change, some positives and negatives of both systems, and cover some lessons learned, as well as tips for others approaching or going through this process.
- ***Glossary project***: S&P had previously worked on a "glossary project" to identify existing digital preservation glossaries. Something to revisit? (last updated November 2022): Glossary Project for Standards and Practice Interest Group

Present (name and institution) (39 total)

- Ann Hanlon (UWM)
- Michael Dulock (CU Boulder)
- Michael Barera (Milwaukee County Historical Society)
- Felicity Dykas (U of Missouri)
- Sibyl Schaefer (UC San Diego)
- Fatemeh Rezaei (University of Baltimore)
- Nathan Tallman (APTrust)
- Adriane Hanson (University of Georgia)
- Peter Gorman (U. Wisconsin–Madison)
- Linda Tadic (Digital Bedrock)
- Carol Kussmann (University of Minnesota Libraries)
- Patrick Daglaris (Oklahoma State University)
- Frances Andreu (Rochester Institute of Technology)
- Allison Look (Illinois Institute of Technology)
- Michelle Paolillo (Cornell University)
- Dina Sokolova (Columbia University)
- David Tenenholtz (RAND)
- Miriam Leigh (Harvard University)

- Jimi Jones (University of Illinois at Urbana-Champaign)
- John Kunze (ARK Alliance | Ronin Institute)
- Karla Roig Blay (UT Austin)
- Hannah Tashjian (UC Berkeley)
- Margaret Turman Kidd (VCU)
- Ima Oduok (Texas Digital Library)
- Andrew Diamond (APTrust)

Minutes

Michael Dulock's slide deck: ndsaSPIG_migratingDigPres_dulock_09jan2024.pptx
A case study of University of Colorado Boulder's migration from Arkivum to AP Trust. Neither an indictment of one nor an endorsement of the other. Arkivum in place 2018. Ended July 2022. Started APTrust in July 2022.

Why change? Biggest relevance was a closer alignment with the library's ingest methods. Looking for less complexity and overhead for packaging and depositing content. Had to open tickets for Arkivum to pull content into their system. Methods of deposit for AP Trust simpler. Both use AWS as their back end.

Cost model mattered. Allocations paid for ahead of time with Arkivum so up to institution to fill it. AP Trust doesn't require pre-paying for space. Pay for space as it's used. The membership model was also attractive because of the community/team effort aspect of it. Not just a client/provider relationship.

Uploaded 7.7TB/374K files by August 2023. Re-secure everything that had been stored in Arkivum. Complications? Export to AWS only: bucket-to-bucket transfer. Export in Arkivum database format. Lots of nested folders. Lots of mapping CSV files to understand what relationships between files were, including replicated files. Needed to use local intermediate storage because folders couldn't be downloaded from original plan, CloudBrowser. Finally, reconciliation revealed that after much checking, some collections were missing.

See slides for Positives and Negatives of both systems, for UC Boulder's purposes.
ndsaSPIG_migratingDigPres_dulock_09jan2024.pptx

Lessons learned: How will your data be served back to you? This might impact costs of migration, how you work with your local IT, etc. Important to maintain a relationship with the provider until the end of the migration. For example, one of the missing collections was finally located in the prep/docking bay and the provider was able to locate it there. Otherwise, that may have been lost. Maintain relationships! Also - overestimate how long migration will take! Document everything - record deposits separate from the system itself. What should you expect to find? Recommends having more than one admin for purposes of continuity.

Questions: How confident about getting stuff back from APTrust in future? Bagit is standard for APTrust, so it's whatever you've bagged and pushed up.

Were there user facing issues with Arkivum, given their fluent focus? UC Boulder was just using it for back end storage. Focus was not on discovery.

AV files/files that are in chunks and running checksums? UC Boulder also AV heavy. Those were the biggest deposits and were saved for last! Didn't have to break any files up happily. No pipeline issues for larger deposits. Importance of interacting with the vendor extensively to understand how it will work.