NDSA Infrastructure Interest Group 2023 Meeting Agendas and Notes

When and Where

- The meetings are held quarterly of the months listed below at 3pm EST:
    - March 20th
    - June 26th
    - September 12th (Tuesday)
    - December 4th
- This Google Doc contains agendas and notes from meetings held in 2023 and can be bookmarked with this link: http://bit.ly/3YG4ePI
- 2022 Meeting notes: https://bit.ly/3ffGuMR
- Meeting recordings are available on YouTube: https://bit.ly/2QRIMmO.
- Join from PC, Mac, Linux, iOS or Android:
    - Meeting ID: **745 482 656**
    - password: DLFzoom
    - Direct Zoom Meeting link
    - US: +1-646-876-9923 or +1-669-900-6833 or +1-408-740-3766
    - Find your local number: https://clirdlf.zoom.us/u/adTIubhwZu
    - International numbers available: https://zoom.us/u/cPpLHpqKX

NDSA Slack

Workspace Link
Join Link

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

NDSA Privacy Policy

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

---

Meeting Notes and agendas follow.

December 4, 2023

Attendance:

- Mark Shelstad (he/him), Colorado State University
- Adriane Hanson (she/her), University of Georgia
- Scott Prater (he/him), University of Wisconsin - Madison
- Kyle Breneman, University of Baltimore
- Hilary Wang (she/her), Brown University
- Ling Meng (University of Wisconsin-Milwaukee)
- Ima Oduok (she/her), Texas Digital Library
- Margaret Turman Kidd (she), Virginia Commonwealth University
- Martha Anderson, (she/her), University of Arkansas
- Bethany Scott (she/her), University of Houston
- Don Gourley, Washington Research Library Consortium
- Este Pope, Dartmouth College
- Dina Sokolova, Columbia University
- Robin Ruggaber, University of Virginia

New attendees:

- Heidi Pettitt (she/her), Loras College
- Mike Gates, Brigham Young University
- Shawn Rounds (she/her), Minnesota Historical Society
- Scott Lawan (he/him), University of Minnesota

Facilitators:

- Robin Ruggaber, Eric Lopatin

Meeting dates for next year:

- Mar 18, 2024 03:00 PM
- Jun 17, 2024 03:00 PM
- Sep 16, 2024 03:00 PM
- Dec 9, 2024 03:00 PM

Agenda:

December's meeting will follow the format of a "solution room" discussion, where attendees can introduce a topic on a specific project or challenge at their organization and gain feedback from the group.

Meeting Notes
Recording posted:
https://www.youtube.com/watch?v=2w0pHUpOEy8

September 12, 2023

Attendance: (35)

- Scott Prater (University of Wisconsin - Madison)
- Nicole Scalessa (Vassar College)
- David Tenenholtz (RAND Corporation)
- Ling Meng (University of Wisconsin-Milwaukee)
- Este Pope (Dartmouth College)
- Linda Tadic (Digital Bedrock)
- Mark Shelstad (Colorado State University)
- Hannah Wang (NARA)
- Kim Gianfrancesco (Vassar College)
- Adriane Hanson (University of Georgia)
- Andrew Diamond (APTrust)
- Deb Verhoff (NYU Libraries)
- Hilary Wang (Brown University)
- Don Gourley (WRLC)
- Sibyl Schaefer (UC San Diego)

New attendees:

- Zeke Crater (UVA)
- Michael Dermody (Syracuse University)
- Kent Gerber (U Minnesota)
- Christina Velazquez Fidler (Bancroft Library, U of Berkeley)
- Ima Oduok (Texas Digital Library)
- Peter Gorman (University of Wisconsin - Madison)
- Catherine Gao (USC Digital Repository)
- Kay Slater (Oak Park Public Library)

Facilitators:

- Eric Lopatin, Robin Ruggaber

Reminders:

- DLF Events, St. Louis:
    - The DLF Forum (#DLFforum): November 13-15
    - Learn@DLF pre-conference workshop day (#LearnAtDLF): November 12
    - NDSA's Digital Preservation 2023 (#DigiPres23): November 15-16
    - Program schedule

Agenda:

- Storage strategies and requirements for research datasets – What makes storage and infrastructure requirements associated with research datasets unique?

    - We welcome Sam Gustman as our guest speaker. Sam is the Associate Dean for Technologies at the USC Libraries, and CTO of the USC Shoah Foundation. He will start us off with a presentation about the USC Digital Repository's approach to providing preservation solutions for researchers that take into consideration the data management requirements of NIH and NSF and other federally funded projects.

June 26, 2023

Attendance:

- Robin Ruggaber, Cal Lee, Michelle Paolillo, Stephen Abrams,Terry Brady, Nathan Tallman, Dan Noonan, Dina Sokolova, Carol Kussmann, Linda Tadic, Leah Prescott, Sean Buckner, Deb Verhoff, Paul Clough, Kyle Breneman, Dianne Dietrich, Kim Gianfrancesco, Margaret Turman Kidd, Mira Basara, Ling Meng

New attendees:

- Rachel Gattermeyer
- Don Richards
- Tyler Thorsted
- Laura Henze
- Mark Cyzyk
- Kevin Latta

Facilitators:

- Eric Lopatin, Robin Ruggaber

Thank you again for adding topics to our annual poll:
https://www.tricider.com/admin/3UqV047LvmN/ERtRmm1CbIV

Agenda

- DLF Events, St. Louis:
    - The DLF Forum (#DLFforum): November 13-15
    - Learn@DLF pre-conference workshop day (#LearnAtDLF): November 12
    - NDSA's Digital Preservation 2023 (#DigiPres23): November 15-16

- Speaking at today's meeting – Please welcome Stephen Abrams, Head of Digital Preservation at Harvard Library, who will present on Harvard's Digital Repository Service (DRS) Futures Project.

Notes

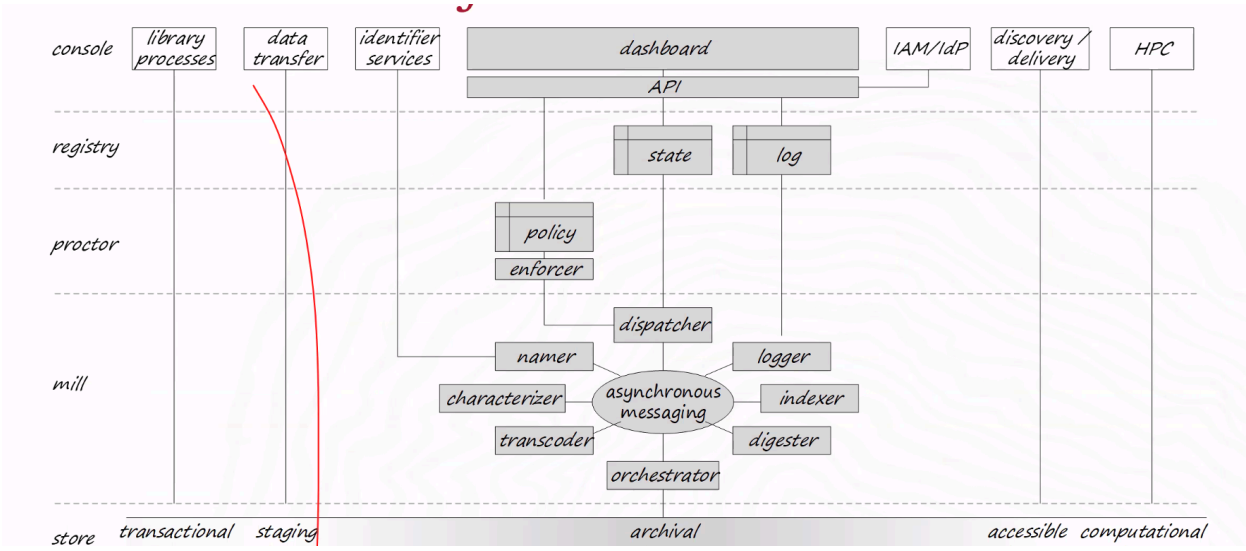
- Revitalizing Preservation Infrastructure: DRS Futures Project @ Harvard
    - Funded through a
- DigiPres @ Harvard
    - Continuation of the Library's historical stewardship mission in the digital realm
    - Initial planning, ~1998
    - Digital Repository Service (DRS) in production, October 2000
    - Remains a custom in-house-developed system - showing its age!! Becoming unsustainable
    - 11M objects, 890M files, 90+ formats, 2 PB
        - Anticipated 2-5x growth over the next few years, many other initiatives happening
- DRS Futures Project
    - Internally funded 3-year project to revitalize core digital preservation infrastructure
        - Phase 1 - Imagine an *ideal* repository (not worrying about doing it)
        - Phase 2 - Plan an *achievable* repository (trim down ideal)
        - Phase 3 - Deploy an *operational* repository
        - See https://sites.harvard.edu/drs-futures/ for details
- Complementary Design Approach
    - Comprehensive functional/non-functional requirements codify the transition from exploratory Phase 1 to pragmatic Phase 2
        - Viewed as a once-in-a-generational activity, 20+ years of life
    - Bottom-up, *inductive* synthesis from literature review, stakeholder engagement, peer consultation
    - Top-down, *abductive* derivation from axiomatic principles
        - inferences, what is the best possible, logical answer, open ended

        - Small set of high-level principles, scaling top-down, adding more detail, subdividing, iteratively refine
    - Done in parallel, overlap in the middle. Merge!
- Conceptual Foundations
    - Digital preservation is not just an exercise in data *management*, but rather, essentially human *communication* cross time

    - Primary imperatives
        - Persistent access to *authentic* information *objects*
        - Persistent modalities of *authoritative* information *performances*
            - Behaviors, implicit or explicit
            - We can't see bits, is the rendered object sufficient?
            - Static or dynamic performance, personalization implications
        - Persistent opportunities for *legitimate* information *experiences*
            - Consumer experience using the material
            - Subjectivity introduce?
            - Does the user trust the object?
    - Watch for our paper iPRES 2023 for more detail.
- Philosophical Foundations

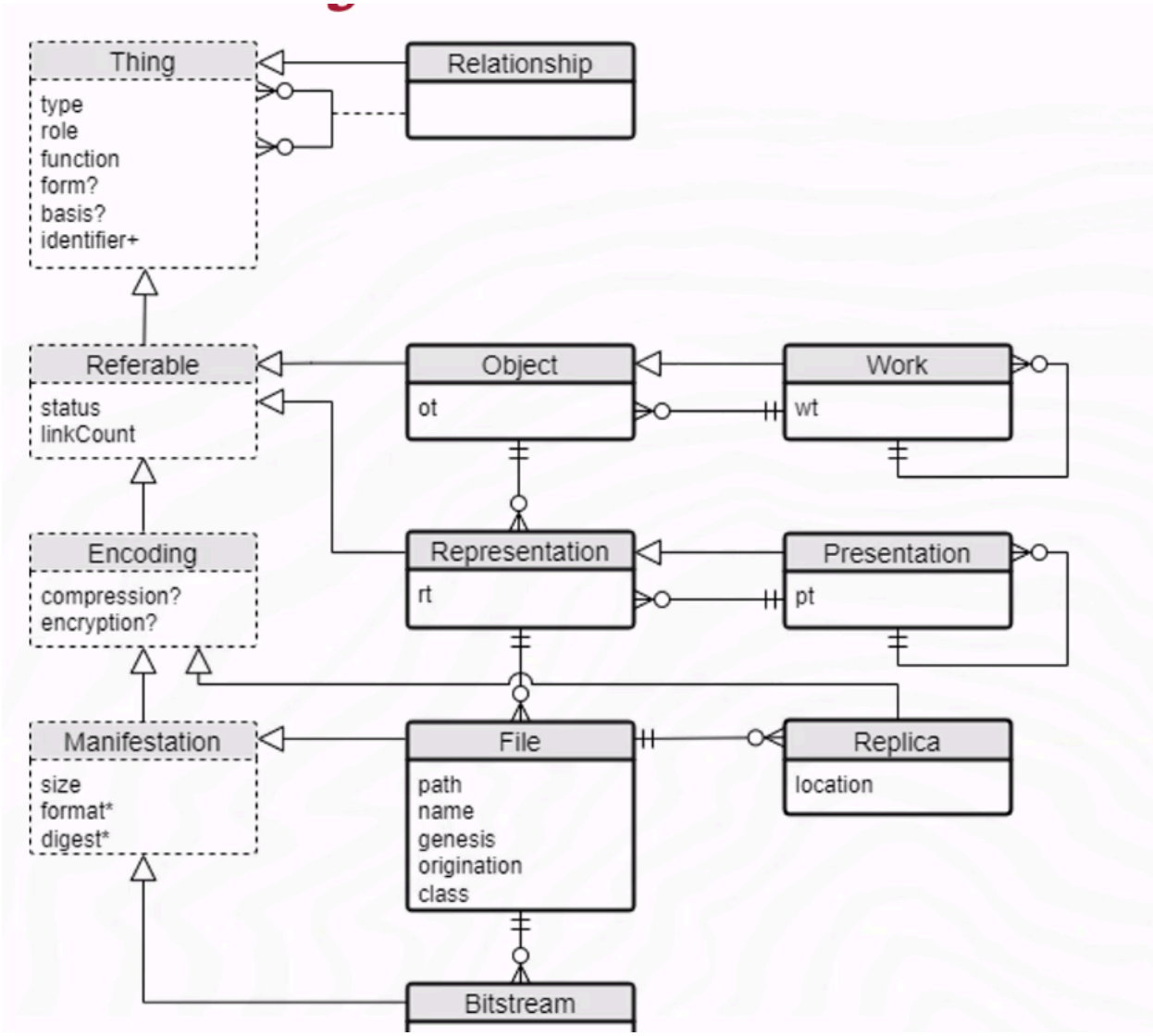
CONCERN	Managerial				Communicative		
REFERENT	Information object				Information experience		
FOCUS	Artifactual				Experiential		
ABSTRACTION	Carrier	Message			Performance	Environment	Mind
FUNCTION	Reifactory	Representational	Rhetorical	Ontological	Epistemological	Associational	Phenomenological
AFFORDANCE	Manifestation	Encoding	Expression	Meaning	Behavior	Context	Understanding
SEMIOTIC	Ontics	Empirics	Syntactics	Semantics	Performics	Plaistics	Pragmatics
IMPERATIVE	Integrity	Validity	Authenticity	Reliability	Accessibility	Relevancy	Legitimacy
DESCRIPTIVENESS	Is-ness		Of-ness			About-ness	
ROLE	Enabling means				Enabled ends		
MEASURE	Quantitative output				Qualitative outcome		
METRIC	Trustworthiness				Success		
EVALUATION	Objective				Subjective		

    - See middle rows
        - Affordance - read left-to-right, builds on itself
- Functional Reference Model



- NOT an architectural diagram
- High-level sense of capabilities and associations that might persist among components
- Darker gray – core infrastructure within a wider ecosystem of library infrastructure
- Repository has 5 layers
  - Console – user facing – patrons and internal managers, GUI, API, CLI
  - Registry – persistence of state information, metadata, state of content, state of system, configuration, operational information
  - Proctor – mode of operation with high-level, policy driven activity, embody human decision making, human knowledge, in machine-actionable way (like iRODS, kind of like PARCORE).
    - Use repo as a state tracker based on any activity. Invoke policies rules based on activity, that may invoke microservices in the mill
  - Mill – microservices, async messaging
  - Store – storage layer, persistence of the content manifestation
- Fault tolerance through asynchronous eventual consistency
- Performance through microservices
- Flexibility though separation of concerns
- Productivity through policy-driven automation
- Enhancement through reconfiguration, not (necessarily) recoding
  - Easier to make changes, e.g. adding support for new format or content genre with minimal coding. Requires machine-actionable language to power an generalized engine.

- Informational (data) Reference Model



- Not as fleshed out right not
- Main hierarchy in middle
  - Full-level, not simple
  - Bitstream is optional! Useful for container files (rep a file in container as a file)
- Left-hands side has parallel abstractions to guide definitions of properties,
  - Thing – everything is a thing with properties
  - Referable – questions of status, e.g. active, deleted, purged. Compose complex representations through direct physical inclusion by value or reference. This will prevent deletion when it's a dependent.
  - Encoding – size, format, characteristic digest
- *Objects* vs *works*

    - *Content* vs *system* objects
        - System – about the system – meta. Aid in file recovery.
    - *Representations* vs *presentations*
        - Similar to METS physical structure and logical structure representations.
        - Distinguish between static structural arrangement and behavioral navigation of the thing.
    - *Digital* vs *tangible*, and *substantive* vs *descriptive* vs *instrumental* representations
        - Tangible – physical object, tangible thing not physically in the repository, metadata only
        - Substantive – core content value, the thing, the information
        - Instrumental – information necessary for proper performance of the thing, E.g., audio playlist, color profile.
    - *Data* vs *metadata*, and *singular* vs *wrapper* vs *container* manifestation
        - Primary date vs metadata data
        - Singular files – whole and complete
        - Wrapper – have internal structure
        - Container files – arbitrary aggregations
- Progress
    - Done
        - Dedicated project roles
        - Project norms and best practices
        - Stakeholder engagement
        - User stories/use cases
    - In Progress
        - FUnctional/non-functional requirements
        - RFP tender
    - To Do
        - Build/buy/integrate decision
        - Procurement/deployment
        - Testing, acceptance, training, productionizing
- Questions
    - Preserving the bits and bytes, @NYU providing digipres for IR and data, research data that we are unfamiliar with that curators will have to manage, where do we get information about what it is that we are preserving? Hang our hat on reproducing the bits. How do you take into account the things we don't know?
        - Harvard also integrates with their IR and data repositories. IR is mostly for open access compliance, so not as likely to have odd formats. The data repository has similarly not present problems, mostly social science data - tabular and numeric.

        - For knowing the unknown
            - Lucky they have deep curatorial expertise in most of the areas with emerging formats or special software.
            - Create bridges in the process, advocate for preservable forms of data.
            - Expanding emulation capabilities through EAASI so they can start collecting software
            - With regard to information model, they want to provide the widest possible opportunity for richest description to enable future activity.
            - If they find something they can't support, try to intervene.
            - Information sharing. Pool and share for widest knowledge
    - Eric's question, missed!
        - Trying not to. Don't want to bias findings.
        - Stephen encourages people to not make assumptions based on current experience.
        - They won't be doing a full build. THey don't think they or anyone can really do that.
            - Tinker around the edges. Value-added functions. Unique insights for something useful that wouldn't otherwise be available.
        - Open. Not sure if it will be a single solution or if they will piece something together in an integrative fashion.
            - If a monolith, one person to call! But plusses and weaknesses.
            - Integration – can be best-in-breed at cost of complexity, distributive maintenance
            - Will to marketplace, OS and commercial.
    - When this is operational, how many dedicated FTEs do you expect to be working on it?
        - Operational – not too large, maybe same as current? No one is full time on it.
        - Maintenance, dev – depends on the solution. Could be fully vended and they prod the vendor and help as they are able.
            - Integration work would become more significant.
        - Hard to say what's reasonable right now. Hope to be able to answer when next phase is complete!
    - It's great you're receiving institutional support for this. Congratulations! Will this include building out Harvard's internal technology infrastructure, to support the storage?
        - For the past 2-years they decoupled storage from the repository application. Part of a larger long-term effort to make library storage available to many library applications/systems/etc. Moving away from

having to move things from A to B between systems. Don't want to drop it on the floor! Let things live where most naturally.
    - THey now lease all storage capacity from internal (research computing IT group) and external providers (regional storage consortia – Northeast Storage Exchange for nearline tape, commercial vendors AWS Glacier Deep Archive for offline equiv, Wasabi for high-performance).
    - Try to keep things where they already are and build things on top. They can do anything through S3 API and use Starfish storage orchestration software for policy-driven replication based on file-role.

March 20, 2023

Attendance:

- Krista Oldham, Martha Anderson, Michelle Paolillo, Paul Clough, Scott Prater, Kim Gianfrancesco, Edson Smith, Dianne Dietrich, Este Pope, Don Gourley, Kyle Breneman

New attendees:

- Miriam Leigh

Facilitators:

- Eric Lopatin, Robin Ruggaber

Thank you again for adding topics to our annual poll:
https://www.tricider.com/admin/3UqV047LvmN/ERtRmm1CbIV

Discussion topics
Martha Anderson – Head of the Mullins Library Digital Services Department at the University of Arkansas – The role of AI now and in the future for digital preservation and appraisal.

Projects and organizations mentioned by Martha:
IDEA Institute on AI: https://idea.infosci.utk.edu/
InterPARES Trust:
http://interparestrust.org/trust/about_research/studies

Off the shelf solutions:

- Aeon: Email archiving
- BitCurator
- Transkribus

Dale Poulter – Director of Library Technology and Digital Services at Vanderbilt University –

3 (copies) -2 (technologies) -1 (local) Preservation strategy. Are multiple AWS Zones adequate or are regions needed?

Links shared in chat:

- DPC AI for Digital Preservation: https://www.dpconline.org/events/previous-events/eventdetail/115/-/ai-for-digital-preservation
- Optical Disc-Based Data Archive System:https://panasonic.net/cns/archiver/
- DNA Storage: https://en.wikipedia.org/wiki/DNA_digital_data_storage
- Filecoin:https://filecoin.io/
- Starling: https://starlingstorage.io/
- Landano: https://www.landano.io/ - Led by Peter Vangarderen, AtoM and Archivematica creator - decentralized preservation and archiving of land rights records
- LOC Designing Storage Architectures (3/27-28):
    - Link to 2022 program and presentations: https://www.digitalpreservation.gov/meetings/storage22.html
    - Link to 2023 program and presentations: https://digitalpreservation.gov/meetings/storage23.html

June 26, 2023

Attendance:

-

New attendees:

-