🌐 Collaborative Good Practices for Digital Preservation in the Cloud

This living document captures insights, strategies, and lessons shared during the "Navigating Digital Preservation in the Cloud" workshop at iPRES 2025. Participants were encouraged to contribute in real-time during each breakout session. Following the workshop, the facilitators edited and refined this document for sharing, utilizing ChatGPT for summarization and synthesis.

💰 Budget, Finance, FinOps

This section of the workshop focused on money: how cloud costs behave, why they're so difficult to predict, and the real-world struggles institutions face when budgeting for digital preservation in the cloud.

The Big Picture

People generally want three things when it comes to cloud budgeting:

1. **Predictability** (to avoid nasty surprises),
2. **Transparency** (to know what they're actually paying for), and
3. **Control** (so they can make intentional choices instead of reactive ones).

Across the board, folks described cloud finances as… complicated. The themes that emerged primarily concerned egress costs, data sovereignty, vendor dynamics, and the varying experiences institutions face, depending on their home country or internal organizational structure.

Key Challenges

- **Egress Charges Are a Maze**
    - The rules and pricing around egress differ from one vendor to another.
    - Uploading your *primary on-prem copy* to the cloud is usually free, unless you use a vendor's physical transfer appliance, which incurs a fee.
    - Moving data *between cloud regions* is a whole different story and can get expensive fast.
    - Most providers offer online calculators, but they're all different, and none feel intuitive or comparable.
- **Data Sovereignty Complicates Everything**
    - Many organizations are required to store their data within specific national borders.
    - That immediately limits where you can store your preservation copies, the number of regions you can use, and which vendor features are even available.
    - This also makes it harder to take advantage of cheaper regions or innovative offerings.
- **Country-Specific Challenges (especially New Zealand)**
    - There are very few data centers available locally in New Zealand.
    - No academic or national cloud provider to serve the research and cultural heritage sector.
    - International tariffs add uncertainty and increase costs.

    - Since most cloud vendors are U.S.-based, support experiences vary widely, ranging from excellent to unhelpful, depending on the representative assigned.
- **Budget Timelines Don't Match Cloud Realities**
    - Many institutions can't commit to long-term contracts because their budgets shift year to year.
    - Storage costs *are expected to decrease over time*, but this is *not guaranteed*.
    - Compute costs change more dramatically and tend to increase when new capabilities emerge.
- **Internal IT Limitations**
    - Some organizations' central IT departments are too focused on security or enterprise-wide responsibilities to support niche cloud preservation needs.
    - That pushes smaller teams into "self-service" models, even if they don't feel fully prepared.

Strategies and Lessons Learned

- **Track Costs Early and Often**
    - Set up cost notifications, anomaly detection, and regular reports.
    - Use historical cost patterns to prioritize upcoming projects (e.g., "this feature will probably spike compute costs").
    - Capture and share cost data with leadership to build understanding and advocacy.
- **Make Use of Reservations (When You Can)**
    - Reserved compute or storage capacity can significantly reduce cost.
    - But this only works if budgets allow for long-term commitments, a challenge for many.
- **Talk in a Shared Language**
    - Finance, leadership, and digital preservation professionals often don't speak the same "cost vocabulary."
    - Using common terminology helps justify budget requests and explain why specific configurations are more expensive (or cost less).
- **Collaborate Up and Down the Chain**
    - There's growing interest in encouraging discussions between cloud vendors and governments about on-shore and off-shore storage requirements.
    - A few participants noted that these conversations can help shape policy and encourage the development of local storage options.

Tools and Approaches Mentioned

- **Containerization** is a method for keeping compute costs lower and scaling efficiently.

- AWS’s serverless **checksum workflow for Glacier tiers**, which temporarily restores objects but claims to be cheaper than doing the whole process yourself after rehydration.
- As **egress rules and pricing vary,** make sure to be aware of the rules and pricing associated with your vendor(s). Costs may be reduced or waived after content has been stored for a specified period.
- Cloud providers sometimes offer **egress cost waivers** for educational institutions.
- Leaning on community knowledge, shared software, and tool registries when internal resources are thin.

Open Questions

- Would the community benefit from a **FinOps group specifically for digital preservation**?
- If so, who would host it — a coalition like DPC, a national body, a consortium, or something more informal?

🔁 Preservation Activities

The Big Picture

This group focused on how core preservation work changes in the cloud, particularly regarding fixity, independence of copies, and format identification. Much of the discussion centered on **trust**:

- **trust** in cloud services,
- **trust** in fixity guarantees,
- **trust** that replicated copies are genuinely independent, and
- **trust** that workflows will scale without exceeding budgets.

People are excited about cloud-native tools but cautious about giving vendors too much control.

Key Challenges

- **Independence of Copies**
  - Preservation copies need to be genuinely independent, not just vendor-managed replicas.
  - Lifecycle-tiered storage (“intelligent” or “lifestyle” buckets) doesn’t necessarily satisfy preservation expectations.
- **Fixity (Especially on Cold/Frozen Storage)**
  - Fixity on deep storage tiers is slow and can be costly.
  - There’s uncertainty surrounding the reliance on cloud and other storage providers’ block-level integrity checks.
  - The frequency and method of fixity checks are significant questions; checks performed by the preservation system (rather than at the storage level) are expensive.
- **Tooling Gaps**
  - DROID doesn’t work well for object storage right now; Siegfried performs better. However, the UK National Archives is currently working to improve DROID's cloud functionality.
  - Keeping signature files up to date and re-running identification tools across a corpus is ongoing work.
- **Upload + Network Issues**
  - Moving data into the cloud is highly dependent on local bandwidth and network conditions.
- **Metadata/Characterization Complexity**
  - Extraction pipelines are complicated to design and maintain.
  - Lambda-based extraction helps, but adds additional moving parts.

Strategies and Lessons Learned

- **Scale with Containers**
    - Containers and horizontal scaling make large-scale fixity checks (PB-scale) much more manageable.
- **Break Workflows into Small Steps**
    - Smaller, independent tasks are easier to containerize and run at scale.
- **Pay Attention to Sovereignty**
    - Region/availability zone choice matters — both legally and technically.
- **Rethinking Fixity**
    - Ongoing fixity is essential, but relying only on your digital preservation system can be an expensive approach. Investigate options to perform fixity checks at other points in your infrastructure, such as the storage system.
    - APTrust's streaming checksum method in containers offers a more cost-effective approach for hot storage (see Notes and Resources).
- **Format Identification in the Cloud**
    - Currently, Siegfried performs better in cloud environments than DROID. [We learned in the Bake Off that there is a version of DROID that will work in a serverless cloud fashion nearing production status.]

Tools and Approaches Mentioned

- Rosetta, Preservica, Dropbox (with caveats).
- Identification: **DROID**, **Siegfried**, **Brunnhilde**.
- Serverless processing: **AWS Lambda** (for conversions and metadata).

Open Questions

- Can organizations rely on cloud vendors' block-level integrity checks?
- How do data sovereignty rules and unstable connectivity affect cloud-only preservation strategies?

🧑‍💻 DevOps Workflows

The Big Picture

This group spent considerable time unpacking what DevOps *means* in the context of digital preservation. A significant theme was that DevOps isn't just about automation or tooling: *it's about understanding workflows, speaking a shared language, and making cloud systems more transparent*. People want to reduce operational pain points while staying grounded in the reasons they initially moved to the cloud. There was a lot of curiosity here, mixed with healthy skepticism.

Key Challenges

- **Uncertainty About DevOps Itself**
  - Many people still aren't sure where DevOps begins or ends.
  - The cloud vocabulary (IaC, pipelines, orchestration, etc.) feels intimidating without context.
- **Cloud Opaqueness**
  - Costs, monitoring, fixity, and performance aren't always easy to see or understand.
  - Teams struggle to get the level of transparency they need to feel confident in cloud workflows.
- **Cloud Reliability**
  - Outages (like a recent AWS glitch) showed how dependent cloud workflows can become.
- **Getting Started**
  - Teams expressed uncertainty about the first step: what to move, why to move it, and how to plan around internal capacity.

Strategies and Lessons Learned

- **Start with "Why the Cloud?"**
  - Before moving anything, ask: *What pain point does this solve?*
  - If the cloud or automation doesn't remove friction, it might not be worth shifting.
- **Learn the Shared Language**
  - Understanding basic DevOps and cloud terms makes conversations with developers and sysadmins easier.
  - People emphasized the value of asking questions, even the most basic ones, to build a shared understanding.
- **Use Cloud as an Opportunity to Rethink Workflows**

    - Cloud transitions are a natural moment to pause, clean up processes, and re-evaluate what's actually needed.

Tools and Approaches Mentioned

- **Automation & Serverless**
    - AWS Lambda is used for serverless automated file conversion and metadata extraction, freeing up staff time.
- **Infrastructure as Code (IaC)**
    - CloudFormation and similar tools make it easier to spin up resources consistently.
- **Planning for Migration**
    - Some organizations are considering moves to Azure and thinking about how to address the same transparency and cost concerns that exist there.
- **Cloud as a Workflow Reset Button**
    - Teams described cloud migration as an opportunity to map workflows and rethink the fixity, ingest, and processing steps from the ground up.

Open Questions

- **The Role of AI**
    - AI may help write scripts or code quickly, but it still needs human oversight.
    - There's interest in how AI could speed up specific DevOps tasks.
- **Data Sovereignty & DevOps**
    - DevOps workflows also need to respect sovereignty requirements, especially for Indigenous communities.
- **Who Does What?**
    - How should DevOps responsibilities be divided between IT teams and preservation teams?
    - Should preservation teams be writing code? Should IT be handling preservation logic? Still unclear.

🧩 Cross-Cutting Themes

Across all three breakout groups, several significant themes consistently emerged. These weren't tied to a single topic; they were shared challenges, shared instincts, and shared hopes for how cloud preservation can (or should) work.

🌍 Data Sovereignty & Trust

- Whether discussing budget, fixity, or DevOps, questions about **where data resides** and **who controls it** kept resurfacing.
- Everyone wants clarity around how cloud vendors handle integrity, durability, and geographic placement.
- Especially important for institutions working with Indigenous communities or government data.

💸 Cost Visibility (or the lack of it)

- Regardless of the topic, cost uncertainty inevitably crept into the conversation.
- Teams want better insight into *ongoing* spending, not just estimates.
- People are worried about surprise bills — egress, compute spikes, cold-storage restores, etc.

🔄 Rethinking Workflows (Not Just Lifting and Shifting)

- Across preservation and DevOps discussions, there was an agreement that moving to the cloud presents an opportunity to **re-examine processes**, rather than simply copying the old ones.
- Questions like: *Do we still need this step? Can this be automated? What's the smallest version of this workflow that still works?*

🧠 Shared Language & Skill Gaps

- Teams need a common vocabulary to work effectively across various roles, including archivists, developers, sysadmins, and finance professionals.
- Many people feel behind on cloud or DevOps concepts, but they're eager to learn.
- Conversations go smoother when everyone understands the "shape" of what's possible.

🛠️ Balancing Automation with Human Judgment

- Automation was celebrated, especially for fixity, metadata extraction, and infrastructure.

- But folks also stressed that automation still needs **human oversight**, especially when interpreting errors or planning long-term storage.

🧩 The Big Takeaway

Everyone is trying to balance **cost, trust, control, performance, and simplicity** — and no one solution fits all institutions.

🌟 Moving Forward: Synthesizing the Open Questions

Across all three groups, people were circling the same kinds of uncertainties: trust, cost visibility, sovereignty, cloud opaqueness, and a general desire for more shared understanding. Even though everyone came in from different angles (finance, preservation, workflows), the forward-looking needs are really similar.

Below is a synthesis of what could help the community move forward, grounded in the questions raised during the workshop.

1. Build Shared Spaces for Ongoing Learning

Across groups, people sought ways to keep the conversations going, particularly around cloud cost behavior, fixed models, DevOps basics, and data sovereignty. A few possible paths forward:

- **Create a recurring discussion group or community forum** (lightweight, not a formal working group) focused on cloud + digital preservation.
- **Collect "micro case studies"** from institutions experimenting with cloud workflows.
- **Build a shared glossary** so teams talk with the same vocabulary.

This would help because much of the confusion across groups stemmed from siloed knowledge.

2. Form a Cross-Community FinOps for Preservation Group

This idea was explicitly mentioned in one group, but it fits neatly across all of them. A FinOps-for-preservation group could:

- Share real cost data and models (even anonymized).
- Discuss sovereignty-driven constraints and tradeoffs.
- Surface "how we actually calculate egress risk" approaches.
- Provide a way to clarify vendor offerings and inconsistencies.

This doesn't need to be heavy or official. Even a quarterly meetup hosted by a consortium or a rotating volunteer team would help.

3. Start Collecting Evidence Around Fixity, Replication, and Cloud Trust

Many of the questions boil down to: *"How much can we trust the cloud to do what it says it does?"* A community-driven effort could focus on:

- Comparing block-level fixity promises across vendors.
- Documenting real-world fixity check behaviors (success, failures, costs).
- Identifying where vendor replication falls short of “independent copy” needs.
- Gathering stories about connectivity issues and cloud-only risks.

This doesn’t require a big project; even a shared spreadsheet or GitHub repo would be a meaningful start.

4. Develop “Starter Playbooks” for New Cloud Adopters

A lot of participants wanted to know:

- Where do I start?
- What should I move first?
- What mistakes should I avoid?

A simple set of starter materials would help new adopters avoid reinventing the wheel:

- A “first 10 cloud questions” checklist
- A “what we wish we’d known” document
- Basic DevOps vocabulary for preservation people
- Budget considerations for leadership

These would lower the barrier to entry and reduce cloud anxiety.

5. Create a Community Testing Ground for Tools

Across the notes, people mentioned:

- DROID is currently not working well in cloud contexts (may soon change)
- Siegfried is currently more effective
- Brunnhilde, Lambda workflows, and various containerized pipelines
- AWS checksum approaches

Instead of individual institutions testing these in isolation, the community could:

- Share benchmarks for cloud-based identification tools
- Compare container strategies for fixity
- Collect patterns for serverless processing
- Test how identification tools handle streaming

A loose “testing club” could save everyone time and help vendors better understand preservation needs.

6. Clarify How DevOps + Preservation Should Work Together

There were real questions about:

- Who should own cloud automation?
- Should archivists be writing code?
- Should IT own preservation logic?

A community-led set of *role expectations* could help:

- "Here's what preservation teams typically manage"
- "Here's what DevOps or IT usually handles"
- "Here's where overlap happens and why"

This helps reduce frustration and set realistic expectations across internal teams.

7. Address Sovereignty and Indigenous Data Concerns More Directly

Sovereignty was a recurring theme in every breakout. There is a real opportunity to build shared guidance around:

- Cloud-region selection under sovereignty rules
- How to document where copies actually live
- What to ask vendors about Indigenous data handling
- How to design workflows that still allow independence and verification

This seems ripe for a focused working group or set of community-developed guidelines.

✨ The Big Opportunity

All of these ideas point to a simple, shared need: **People want community, clarity, and confidence.** Not perfect answers, just spaces to learn together, compare notes, and reduce uncertainty. The good news is that none of these needs require massive infrastructure or funding.

Most can be tackled with:

- a shared document
- a standing call
- a GitHub repo
- a small volunteer rotation
- or an informal community-of-practice model

References

📎 Notes and Resources

- Brady, T., Lopatin, E., and Strong, M (2020). Streamlining Merritt Microservice Configuration. https://uc3.cdlib.org/2020/11/05/streamlining-merritt-microservice-configuration
- Digital Preservation Coalition. (2017). Digital Preservation Handbook: Cloud Services. https://www.dpconline.org/handbook/technical-solutions-and-tools/cloud-services
- Diamond, A. (2024). Running Fixity Checks in the cloud. *APTrust News*. https://aptrust.org/2024/01/24/running-fixity-checks-in-the-cloud/
- FinOps Foundation
- Rosenthal, D. (2020). Archival Cloud Storage Pricing. https://blog.dshr.org/2020/03/archival-cloud-storage-pricing.html
- Rosenthal, D. (2019). Cloud for Preservation. https://blog.dshr.org/2019/02/cloud-for-preservation.html
- Ruffner, A. (2024). Simplify to Amplify: Rearchitecting for Preservation in the Cloud. iPres 2024, Ghent, Belgium. Zenodo. https://doi.org/10.5281/zenodo.13226918
- Tallman, N. (2021). A 21st Century Technical Infrastructure for Digital Preservation. *Information Technology and Libraries*, 40(4). https://doi.org/10.6017/ital.v40i4.13355
- Tallman, N. and Diamond, A. (2024). How We Reduced Our Energy Consumption by Switching from Ruby on Rails to Go. *APTrust News*. https://aptrust.org/2024/03/06/how-we-reduced-our-energy-consumption-by-switching-from-ruby-on-rails-to-go/

Facilitator Contacts

- Nathan Tallman, Academic Preservation Trust, Nathan.Tallman@aptrust.org
- Flavia Ruffner, Academic Preservation Trust, Flavia.Ruffner@aptrust.org
- Eric Lopatin, California Digital Library, Eric.Lopatin@ucop.edu
- Terry Brady, California Digital Library, Terrence.Brady@ucop.edu

Cloud Terminology

🧱 Core Cloud Models

Term	Definition	Example
IaaS (Infrastructure as a Service)	Rent servers, storage, and networking; you manage the software.	AWS EC2, Google Compute Engine

PaaS (Platform as a Service)	Build and deploy apps without managing servers.	AWS Elastic Beanstalk
SaaS (Software as a Service)	Use cloud-hosted software through your browser.	Google Drive, Microsoft 365

⚙️ Cloud Operations

Term	Definition	Example
DevOps	Combines development and operations to automate builds, testing, and deployment.	Continuous Integration (CI/CD) pipelines
FinOps	Managing and understanding cloud costs; balancing cost, performance, and value.	AWS Cost Explorer
Container	A self-contained package that runs the same anywhere.	Docker, Kubernetes
Serverless	Run code without managing servers—the cloud scales automatically.	AWS Lambda
Virtual Machine (VM)	A “computer inside a computer” running on shared hardware.	AWS EC2
Infrastructure as Code (IaC)	Enables users to model and manage infrastructure resources.	AWS CloudFormation

☁️ AWS & Cloud Provider Terms

Term	Definition
S3 (Simple Storage Service)	Object storage for files and metadata—common for preservation.
EFS (Elastic File System)	A scalable, elastic cloud storage service that is compatible with the Network File System (NFS) protocol.
EC2 (Elastic Cloud Compute)	A virtual machine that provides scalable, resizable compute capacity in the cloud.

RDS (Relational Database Service)	A virtual machine for running a relational database engine like MySQL or PostgreSQL.
RI (Reserved Instance)	A pricing option that provides a significant discount on services (e.g. EC2, RDS) by committing to a 1- or 3-year usage plan for specific instance configurations.
RCS (Reserved Capacity Storage)	A pricing option that allows for a discount by pre-purchasing an allocation of cloud storage via a 1- or 3-year contract.
ECS (Elastic Container Service) / Fargate	Run containers without managing servers.
IAM (Identity and Access Management)	Controls user permissions and access.
ARN (Amazon Resource Name)	Identifies a resource unambiguously across all of AWS.
Region / Availability Zone	Data center locations that improve redundancy.
Egress	Moving data *out* of the cloud; often incurs costs.

🧾 Digital Preservation Terms

Term	Definition
Fixity	Ensures a file hasn’t changed over time.
Checksum	A digital fingerprint used to verify fixity.
Characterization	Identifies file format and properties.
Replication	Storing multiple copies for safety.
Authenticity	Confirms a digital object is genuine and unaltered.

🧭 Additional Terms

- **Cloud Provider** – A company that offers cloud services, like Amazon Web Services (AWS), Google Cloud, or Microsoft Azure.
- **Region / Availability Zone** – Physical locations of data centers; choosing multiple zones increases reliability.
- **API (Application Programming Interface)** – A way for software programs to talk to each other automatically.
- **Erasure Coding** – A data protection method that breaks data into fragments, adds redundant "parity" fragments using mathematical formulas, and distributes them across different storage locations.
- **Object Storage** – A way of storing data as discrete “objects,” each with its own metadata, rather than in folders.