iPRES 2024, 17 Sept. - Kraakhuis

Collaborative notes by

- Joshua Ng
- Matthias Töwe
- Please add your name here
- Please add your name here
- Please add your name here
- Please add your name here

Implications of Cloud Adoption for Digital Preservation, 11:00-12:30

File Fixity in the Cloud: Policy, Business, and Technical Considerations
Notes:
Initially managed in-house, then hit 100TB mark, worried about managing in-house. Since 2019, fully in the cloud. Storage and web app and everything, in the cloud. AWS.

Medusa. Built it as a file system. Native environment of e-records.

Promise of “Lift and shift” — $540,000 per year!!!!! AWS Elastic File System. This was when it was at 100TB.

First slap in the face. Everything in the cloud is metered.
Move away from hierarchical file system into Flat Object File System. Adopting Object Storage and S3 intelligent tiering. Intelligent tiering, the more you access, it would be in more expensive storage tier. Touch it less, it’ll be moved to less expensive storage tier. 8% accessed frequently. File path as S3 key in object storage.

Had to rewrite fixity checking process. Looked at Lambda, but too expensive, $20,000+. End up using EC2, cost about $1700.

New technical policy –

1. Run two fixity checks on all new items in primary storage
2. Check fixity continuously on a modest pace

Comment:
TU Vienna moved back from cloud to internal storage due to concerns about keeping control of essential university heritage. Similar concerns here?

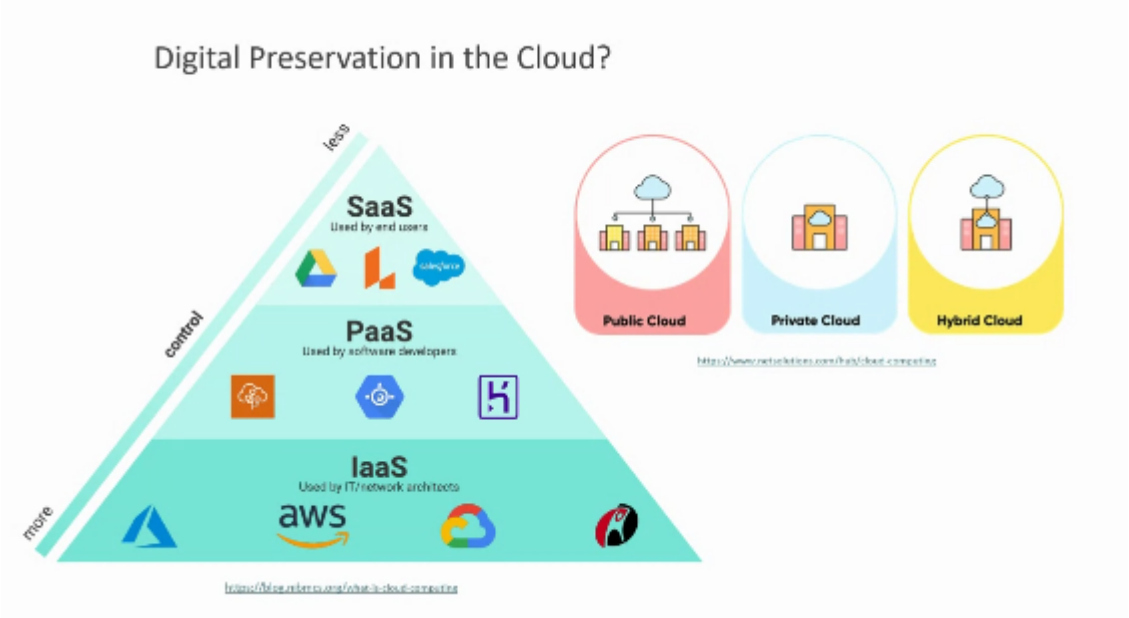
Have fees increased since the migration?

Difficult to keep track of a lot of different bills, therefore need tech people to look into financial side, too.

Is there an exit strategy, e.g., for changing to a different vendor?
The more of their proprietary technology you use, the more you get locked in. So trying to avoid that.
We don't want to be locked in.

Didn't full fixity check compromise intelligent tiering?
No, check was done on the glacier copy. Otherwise, there would be an issue.

Digital preservation in the cloud? Panel
Notes:



Eld Zierau, NL of Denmark
Concerns in the cloud:
Confidentiality issues: use encryption
Transparency: software stack not known and comparisons and auditing not possible
Some non-confidential data are in the cloud; note, however, that metadata may also be confidential in some cases

Jeffrey van der Hooeven, KBNL
2 billion files. Our own IT infra. Moved to private cloud. Hesitant to jump fully into full cloud solution.
We have responsibility. Don't trust data warehouse to preserve it for us.
MS Azure cloud. Lowest tier. Tape machine. In the end it's just tape again :D :D. S3 interface on top of it.

Steve Daly, TNA UK

Lots of effort went into supporting/updating legacy soft- and hardware (which was cutting edge at some point); rather invest this into innovation.
UK gov has cloud-first approach. TNA seized that. We went to the top of the triangle - SaaS.

Micky Lindlar, TIB Hanover Germany
On-prem. Reminds them of 1995 digipres discussion - authenticity, energy cost.
Data privacy and data protection play a big role. Even with on-prem, we had lots of conversation with data protection folks.
Federal state. Always fight for funding.
Universities insist on their autonomy and run own infrastructure: pressure for on-prem first.
Institutional IT looks after the whole institution, compete with DigiPres needs.
We have specific requirement. We want full transparency, also with respect to certification..
Difficult. Pro-SaaS.

Andrea Goethals, Manager, Digital Preservation and Data Capability. National Library of New Zealand
On the trajectory to move digital preservation application and permanent storage to cloud.
Context in NZ: Cloud-first policy. Reality is constrained by limited local options and data sovereignty concerns. Later this year, MS opening data center, based in Auckland (8 hours away), one geographical location. Other providers have data centers in Australia - some departments turn there, but complex for compliance and some material must stay in the country. Keeping data in NZ is vital.
Moving to the hyperscale cloud from private cloud, we know we will lose control. Prefer Cloud-Smart policy.
Key component in private cloud expiring in 1.5 year.
Too expensive, can't afford it anymore
Originally we wanted to move only preservation-storage. Right now, we want to move preservation storage and digipres application, performance reason. Too much latency, so we want to move them into the same place, together.
Rosetta moving to the cloud.
Financial - we want to understand how the cost work. Guardrails in place.
Lots of opportunities. Done the calculation, cheaper for us to make this move. Especially if it's over a long period. Rethink digipres function. Fixity (ref Kyle), take advantage of in-built tools. Research and analysis tool - able to do this if in the cloud.
Prof development for our team.
Long standing IT consultant. We worked with long time. They're helping us make the right decision, to also change our workflow to use the cloud infra.

Questions
How about working with other institutions or in public cloud infrastructures (e.g., SURF in NL)? This is common for research data.
TIB talking to partner for hosting third copy.
Issues can be with the level of security the respective apps offer.
KB NL: has already moved to a government center to share some storage facilities/ not yet moved to the cloud but will in due future (Rosetta preservation system).

Keeping data portable over their lifetime and having an exit strategy is vital. There may be several migrations between solutions over that lifetime depending on how solutions evolve.

Are we in danger of losing skills and capabilities in digipres if we outsource vital functions? Not necessarily - valid approach to outsource “heavy-lifting” which does not really distinguish digipres form other services.

Steve: After we moved to SaaS/Managed service, digital archivist, developers can finally spend time doing digital preservation, things we wanted to do but couldn’t, because we were just feeding and watering the computer processes.

Declouding - describe moving from cloud to cloud, cloud to on-prem.
Transclouding?

Steve, TNA:
Control of the record. We hold our own copy in a cold dark room - custodial copy - in our own premise. Long term migration, Disaster recovery. OCFL. We harvested all metadata and store that as well. Totally decoupled from live system. Custodial copy in data tape system. It’s open source. Anyone can use it.
We don’t let any of our system talk directly to the cloud vendor. We have an intermediate layer. So that we can easy to swap in swap out the different components.

Andrea: Key thing in the next 5 years. Education. How to get application to the cloud to work for us and not against us. Tough for institution without IT staff. Opportunity for us to build a community network of institution who have done this to share our experience and mistakes, lessons learned.
Case example: Developer turned on logging. Received a surprise bill. Went back to AWS, the bill reduced a bit but not by much. Big lesson learned.

Micky: we put things into the cloud uncompressed but we don't know if and how they are compressed there