Implementing Infrastructure

Steps To an Ecology of Digital Preservation

Peter Anderton
Justin Simpson

Shared Service and Common Purpose: Digital Preservation as Infrastructure
DPC/Jisc Event 20 March 2018

Who We Are

Peter Anderton

Product Director
Preservica

Justin Simpson

Director of Technical Services,
Archivematica
Artefactual Systems Inc.

Slides available at: https://bit.ly/DPCDPI

Digital Preservation

Goal

To develop a new national shared technology service.

Purpose

To design, procure and test a prototype RDSS on which a business case will be built for implementation of a full shared service, so that researchers are able to deposit finalised research objects for publication, discovery, safe storage, long-term archiving, reporting and preservation platforms and tools.

Vision

Research Data Shared Service enables open science through efficient and effective capture, preservation and reuse of research data.

Digital Preservation

Vision

Enabling open science through efficient and effective capture, preservation and reuse of research data.

Purpose

Researchers are able to deposit finalised research objects

for publication,

discovery,

safe storage, long-term archiving, reporting and

preservation platforms and tools.

Digital Preservation

Goal

To develop a new national shared technology service.

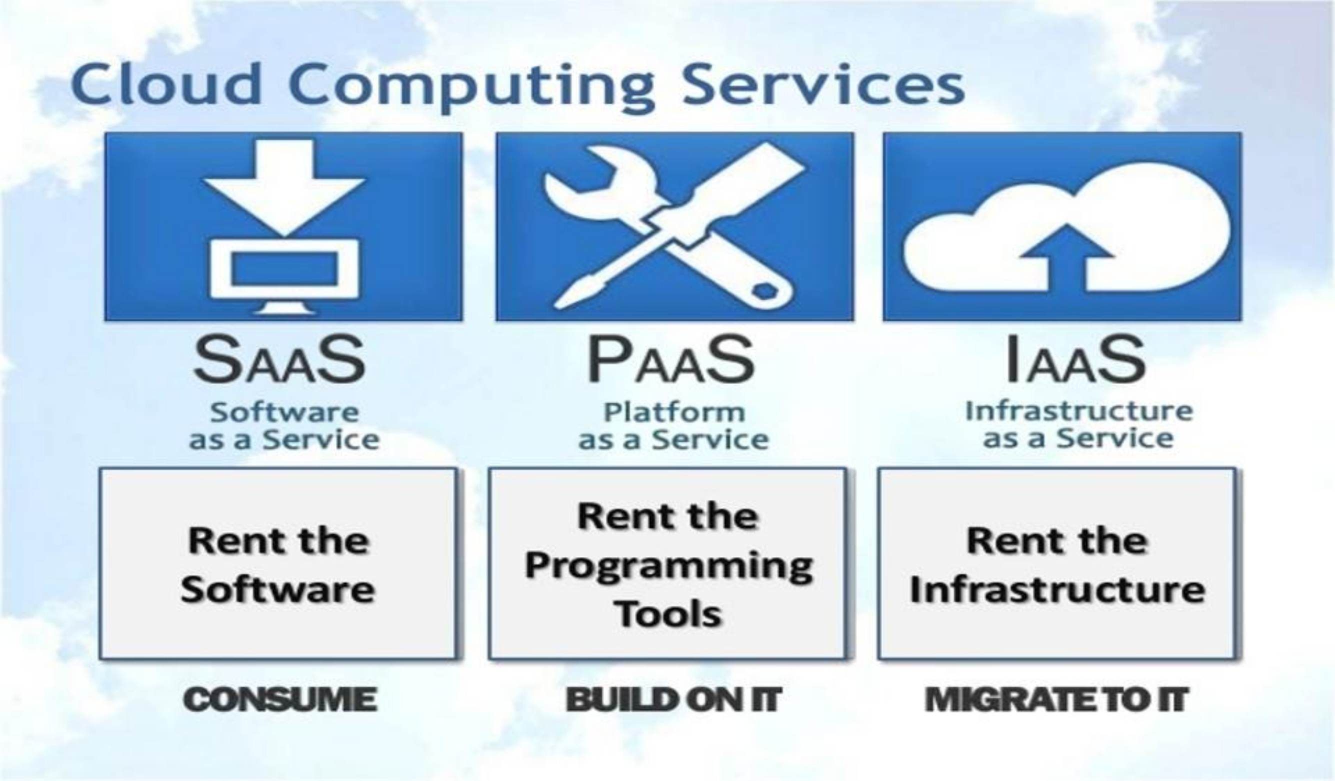
Purpose

To design, procure and test a prototype RDSS on which a business case will be built for implementation of a full shared service, so that researchers are able to deposit finalised research objects for publication, discovery, safe storage, long-term archiving, reporting and preservation platforms and tools.

Vision

Research Data Shared Service enables open science through efficient and effective capture, preservation and reuse of research data.

Digital Preservation



Cloud Computing Services

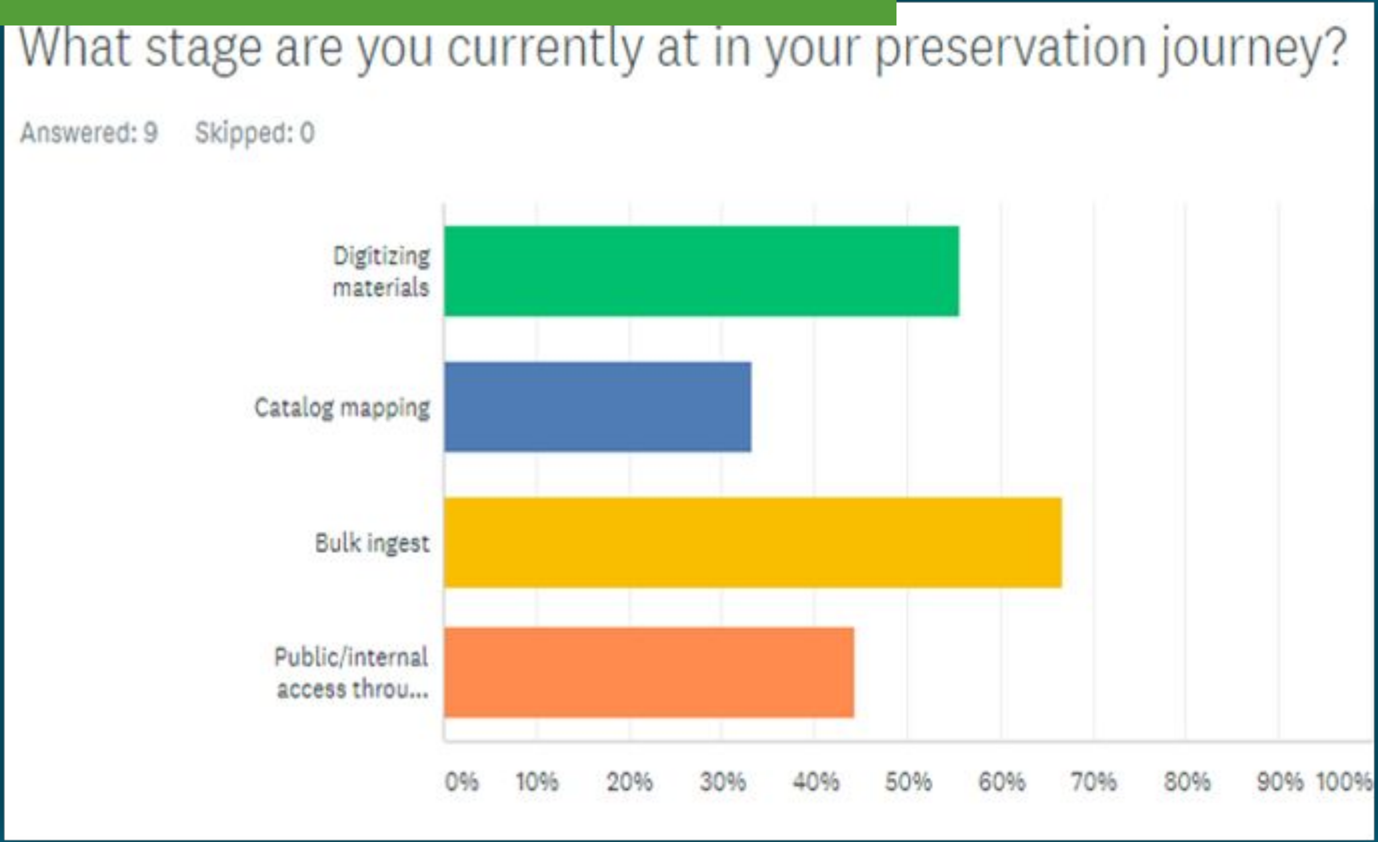
SAAS	PAAS	IAAS
Software as a Service	Platform as a Service	Infrastructure as a Service
Rent the Software	Rent the Programming Tools	Rent the Infrastructure
CONSUME	BUILD ON IT	MIGRATE TO IT

Digital Preservation

Sharing is More than technology

- Awareness
- Involvement & Ownership
- Enablement
- Support

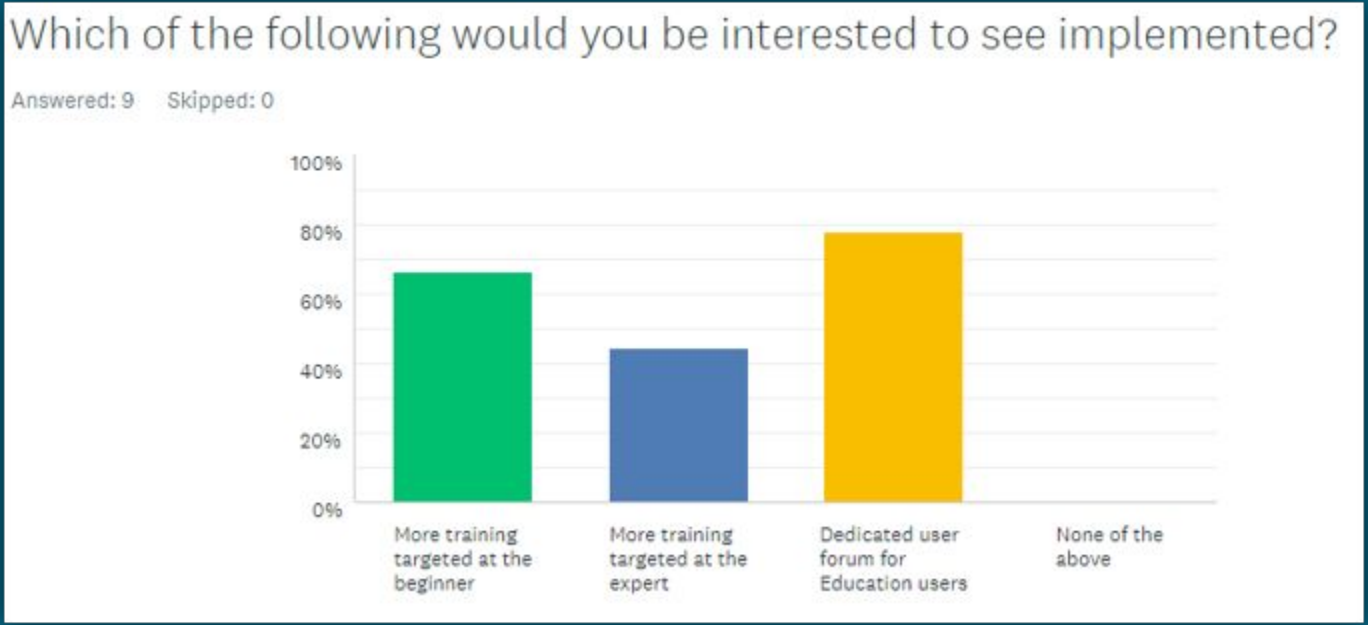
Digital Preservation



What stage are you currently at in your preservation journey?

Answered: 9 Skipped: 0

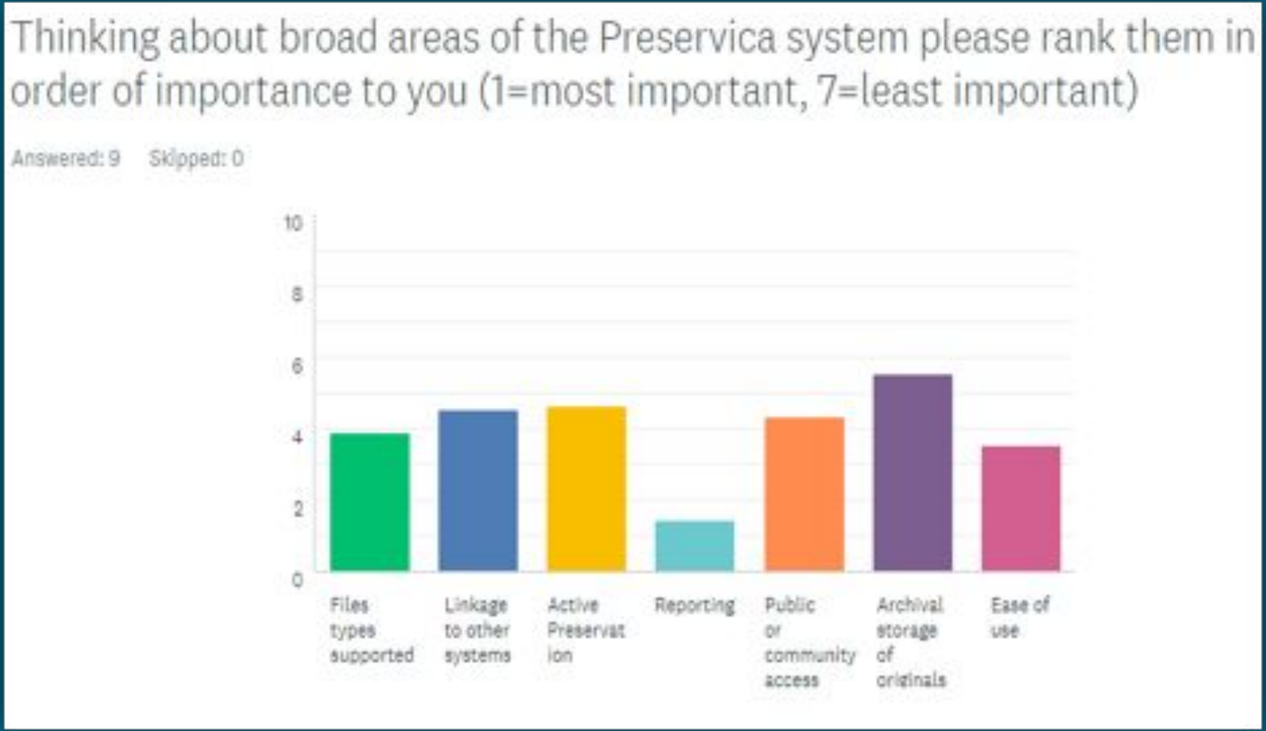
Digital Preservation



Which of the following would you be interested to see implemented?

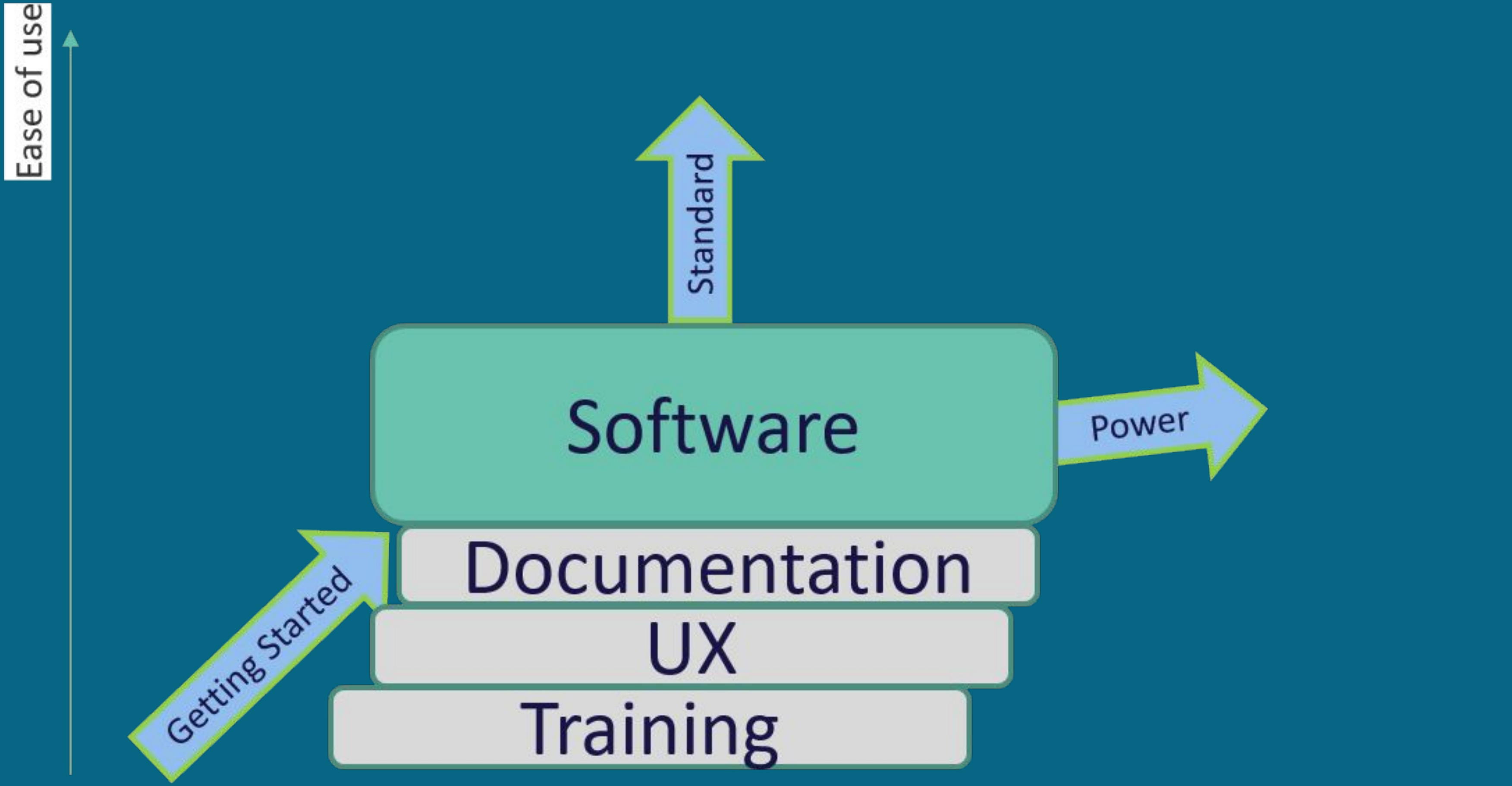
Answered: 9 Skipped: 0

Digital Preservation



Thinking about broad areas of the Preservica system please rank them in order of importance to you (1=most important, 7=least important)

Answered: 9 Skipped: 0


Ease of use
Standard
Software
Power
Getting Started
Documentation
UX
Training
Power

Digital Preservation

Sharing is about...

- Awareness
- Involvement & Ownership
- Enablement
- Support & Community
- Technology

Shared Service Models

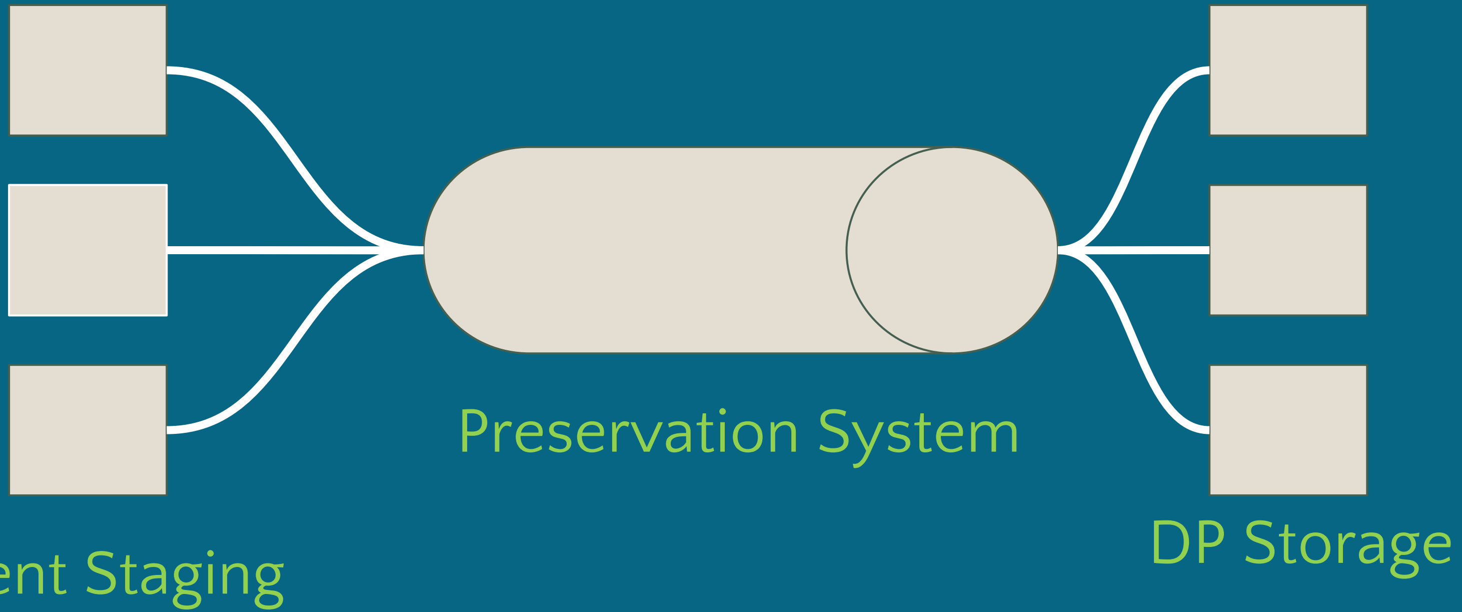
A Shared Service requires placing some elements of the system under some form of centralized control.

Which bits are shared and what form of centralization?

There are many ways to share . . .

Shared Pipeline

Processing occurs in a single shared set of compute resources while separate storage areas are provided for each institution.



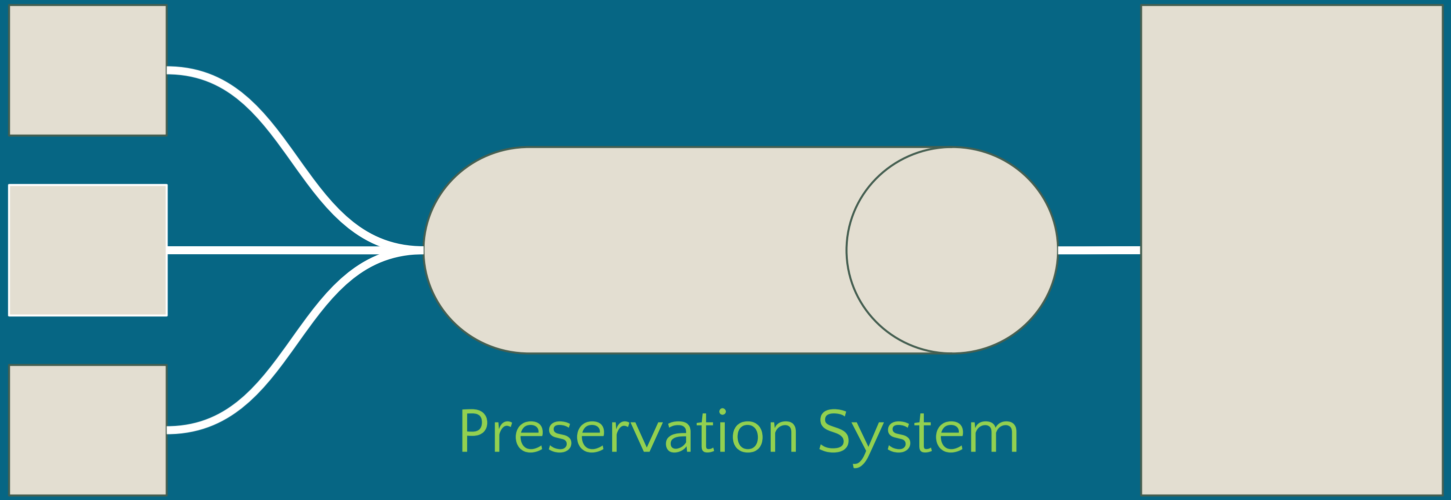
Content Staging

Preservation System

DP Storage

Shared Pipeline

Digital Preservation Storage may be managed by the shared service.



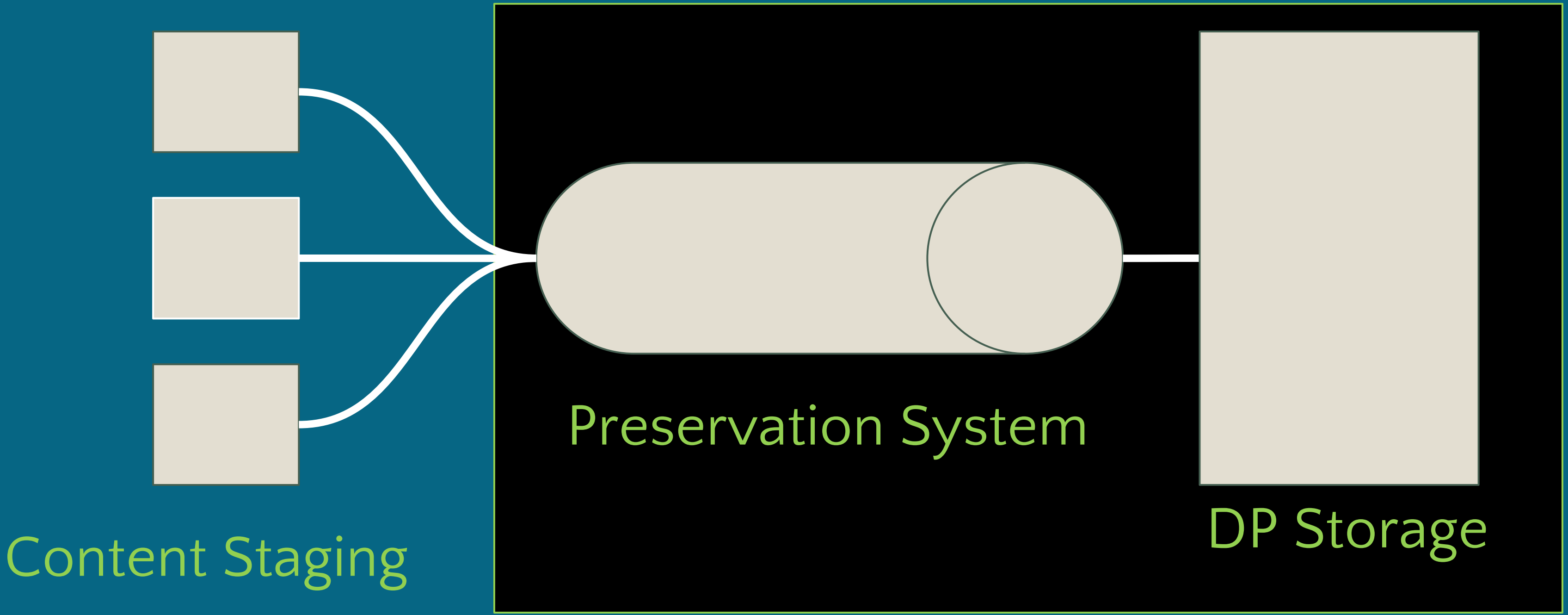
Content Staging

Preservation System

DP Storage

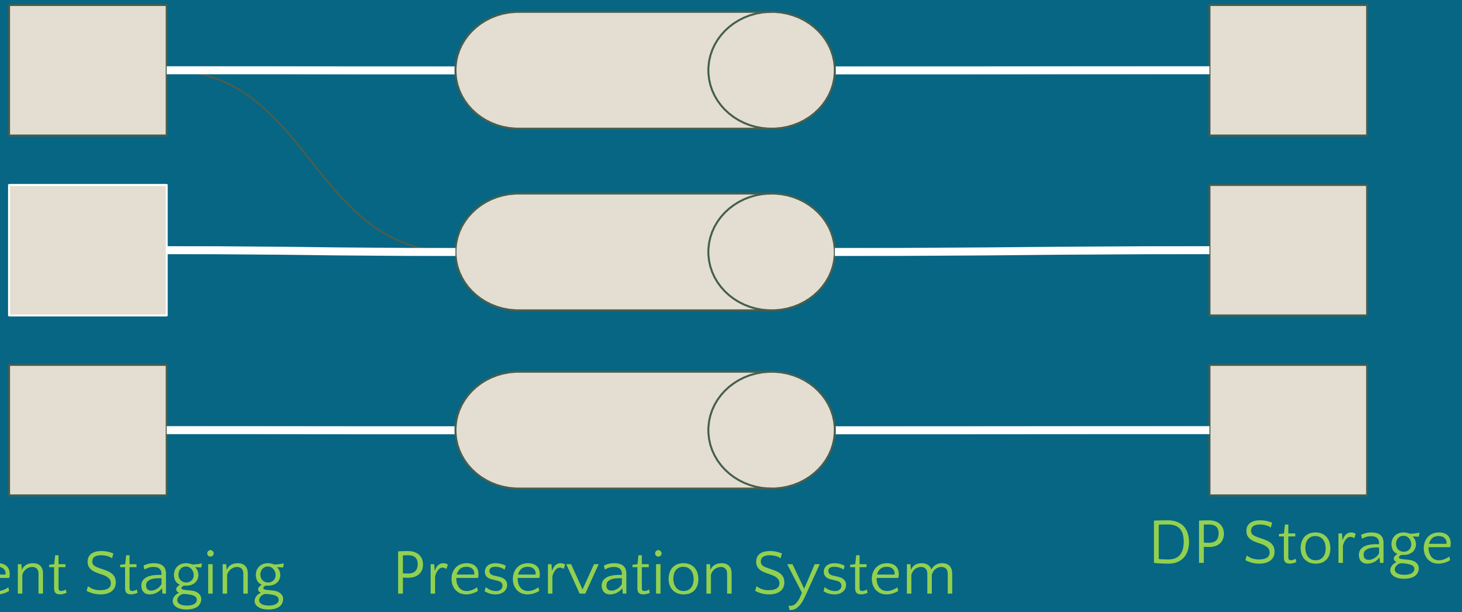
Closed Pipeline

Preservation System may be run as a black box by the Service Provider.
Organisations using the service do not have access other than content upload.



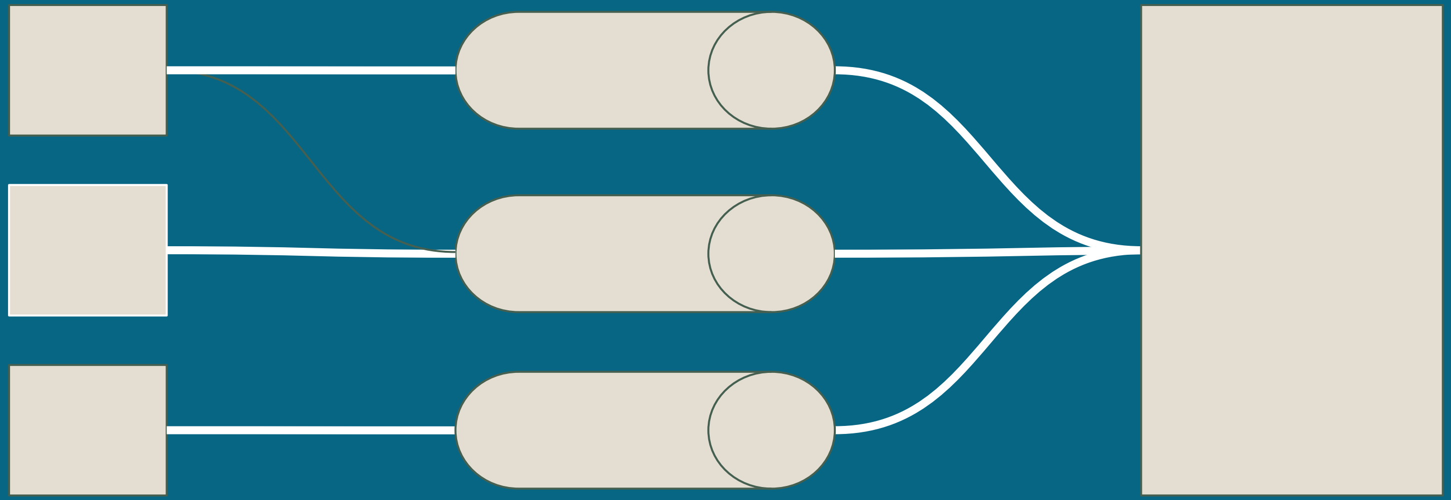
Dedicated Pipelines

Each user is allocated its own discrete computing environment. The underlying compute platform is common, deployment/maintenance is managed centrally.

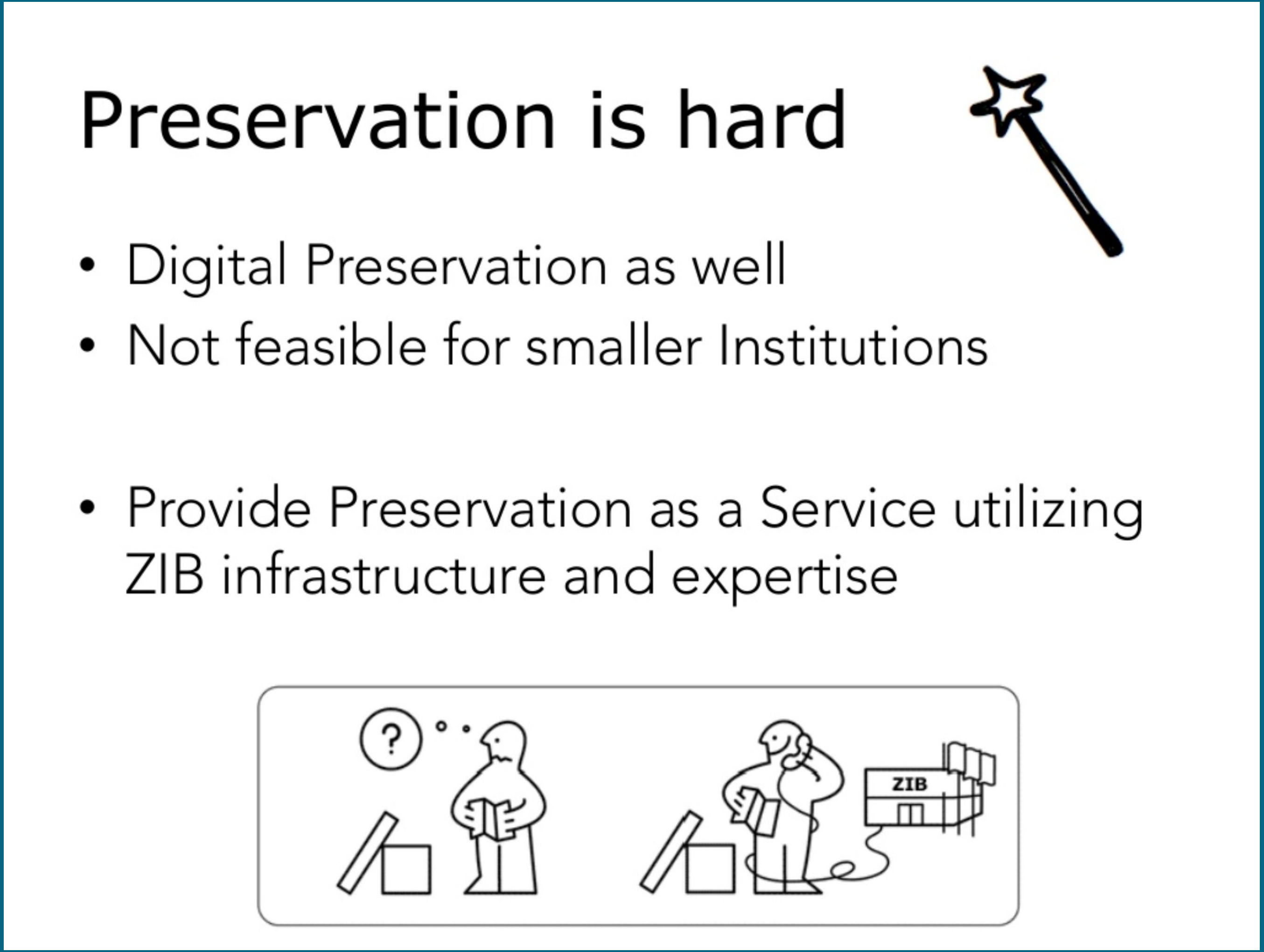


Dedicated Pipelines

Digital Preservation Storage may be managed by the shared service.



Zuse



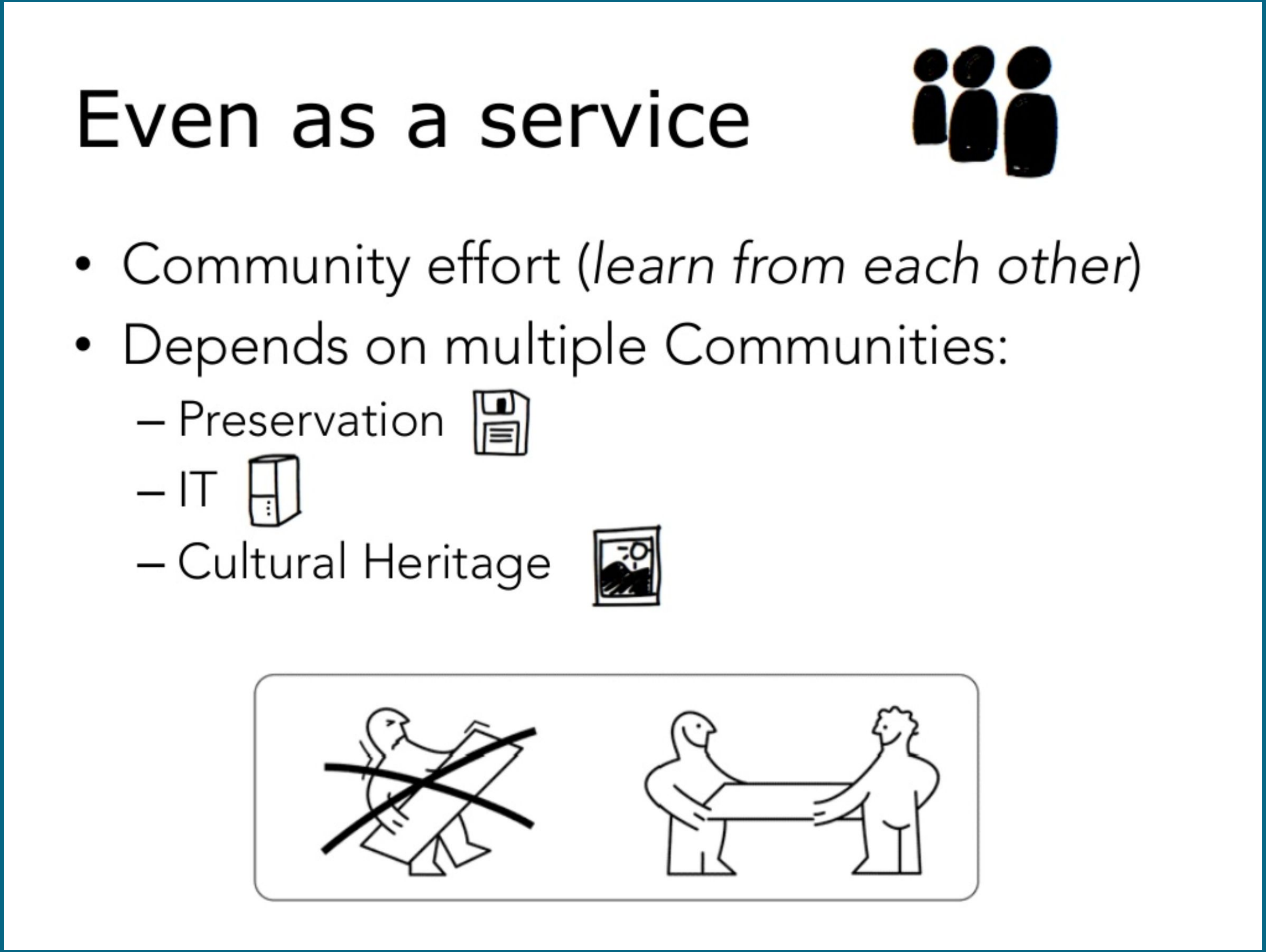
Preservation is hard

- Digital Preservation as well
- Not feasible for smaller Institutions

- Provide Preservation as a Service utilizing ZIB infrastructure and expertise

Zuse Institute Berlin
Marco Klindt and
Kilian Amrhein, 2015

Zuse



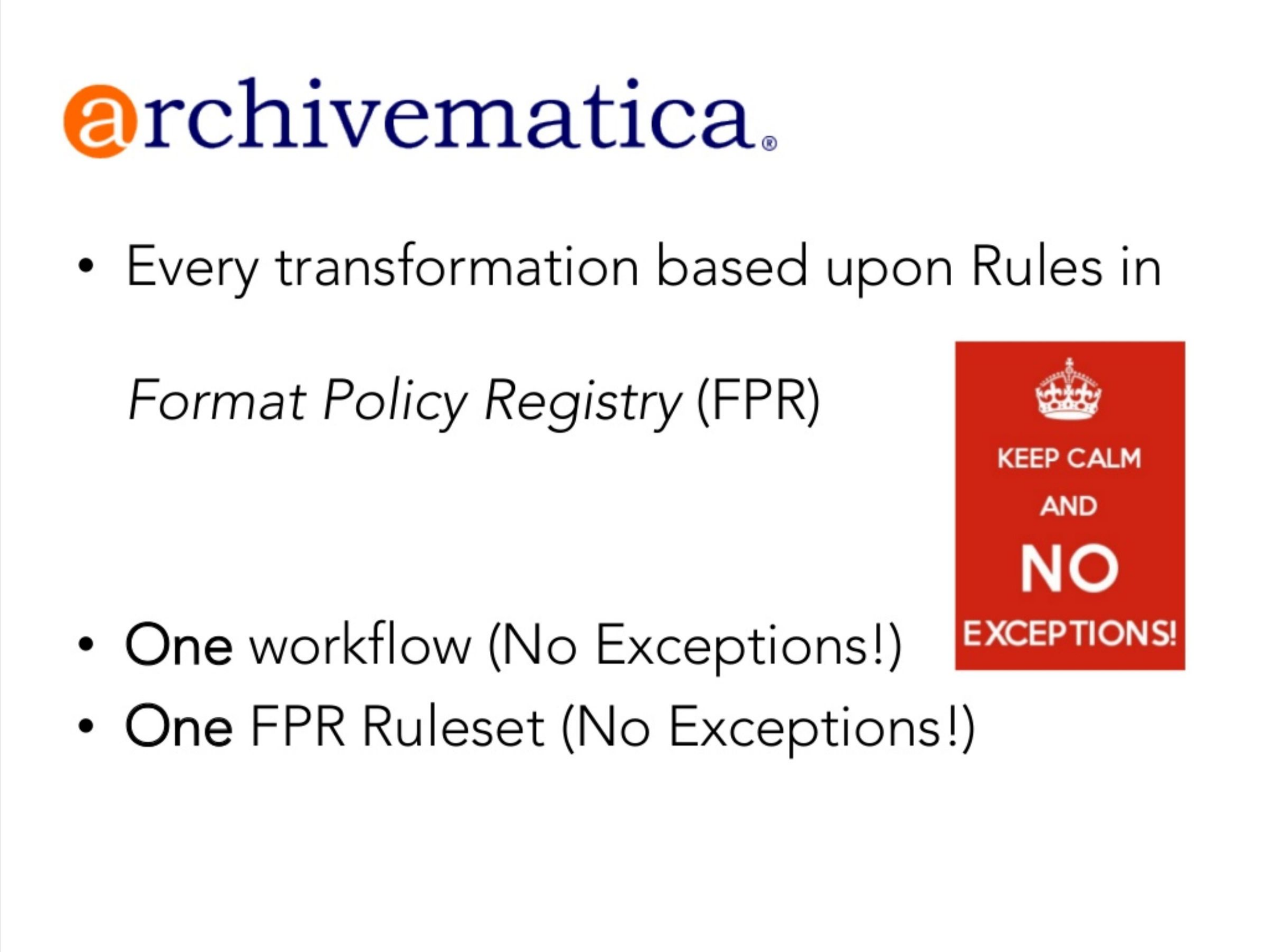
Even as a service

- Community effort (*learn from each other*)
- Depends on multiple Communities:
  - Preservation
  - IT
  - Cultural Heritage

Zuse Institute Berlin

Marco Klindt and
Kilian Amrhein, 2015

Zuse



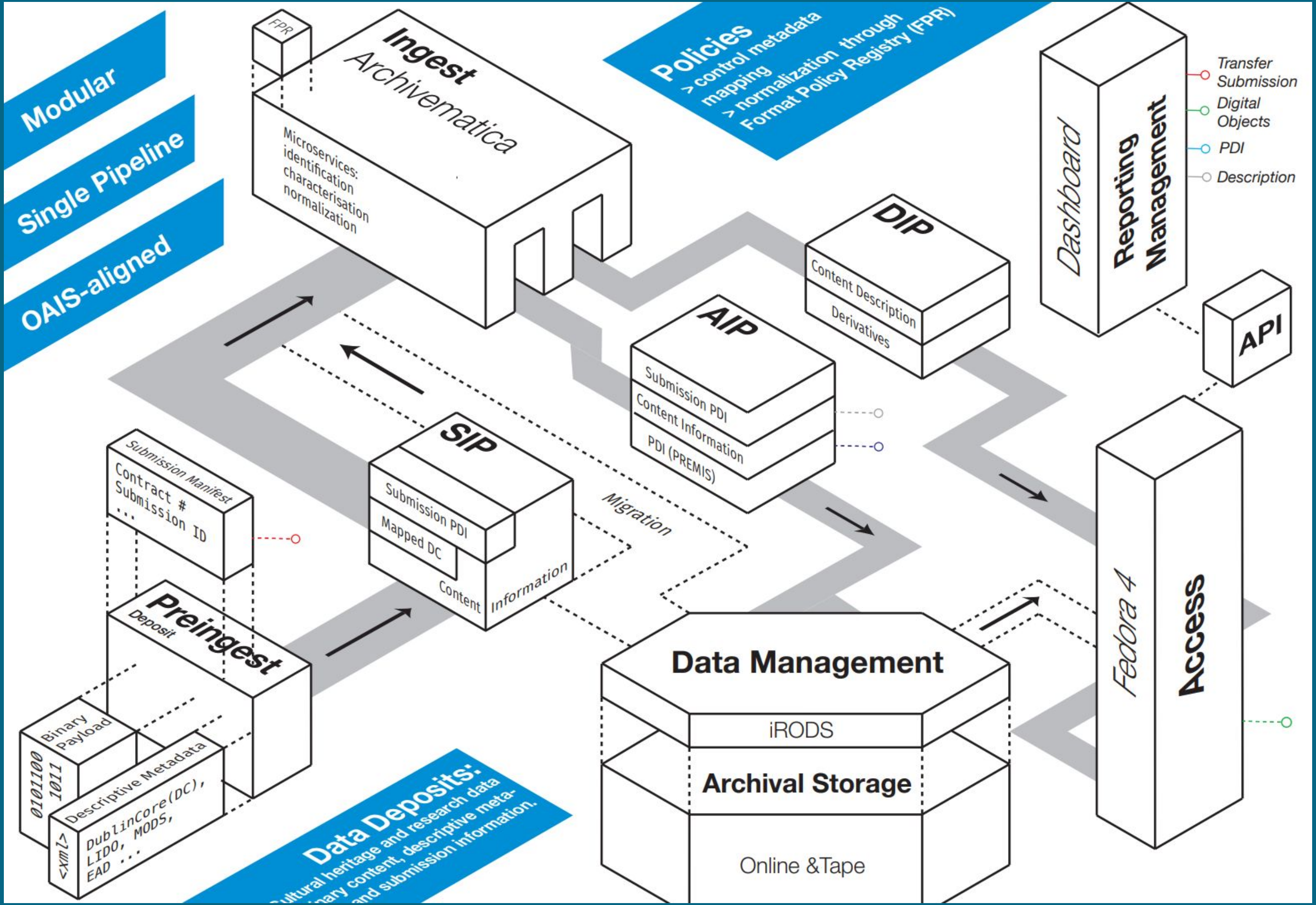
archivematica®

- Every transformation based upon Rules in *Format Policy Registry* (FPR)
- **One** workflow (No Exceptions!)
- **One** FPR Ruleset (No Exceptions!)

KEEP CALM
AND
NO
EXCEPTIONS!

Zuse Institute Berlin
Marco Klindt and
Kilian Amrhein, 2015

Zuse



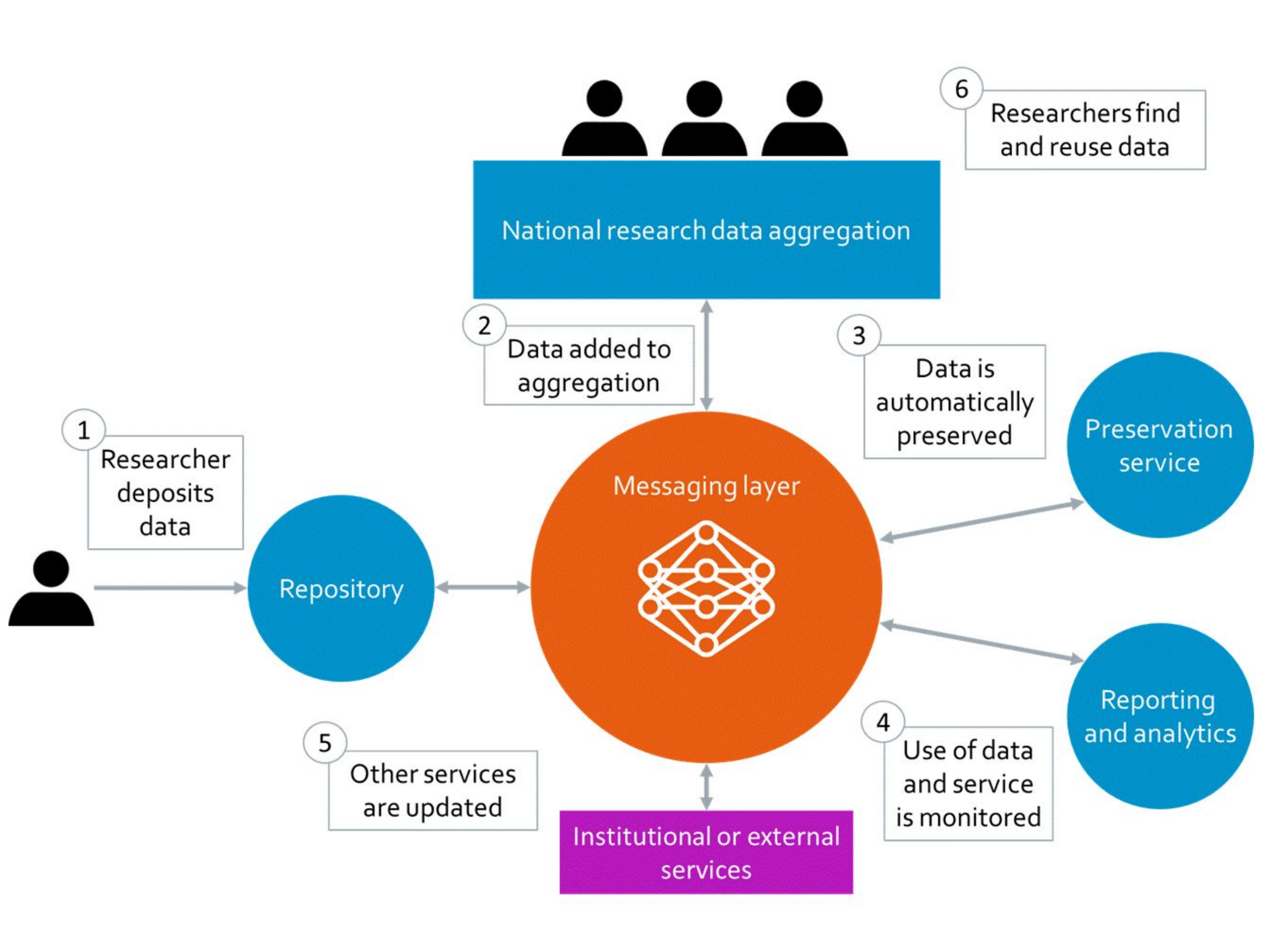
Zuse Institute Berlin

2018 Diagram

Jisc

Preservation System is only one component of RDSS.

Preservation System can be treated as a black box by researchers, and as an open system by Library/RDM staff.


6
Researchers find and reuse data
National research data aggregation
2
Data added to aggregation
3
Data is automatically preserved
Preservation service
1
Researcher deposits data
Repository
Messaging layer
Reporting and analytics
4
Use of data and service is monitored
5
Other services are updated
Institutional or external services

Commercial Shared Services

Commercial Providers that offer cloud based services use a shared services model, in addition to offering on premise / hybrid solutions.

Sharing can happen even with the organisations do not realize or care that they are sharing.

	Preservica Cloud Edition	Arkivum Perpetua	Artefactual Hosted Services
User Login	open	open	open
Resources	dedicated/shared	dedicated	dedicated
Platform vendor	AWS	AWS	OVH
Distributed storage	AWS	Arkivum	Azure
Access	Preservica/Local	AtoM/Local	AtoM/Local

Consortia Shared Services

Service	COPPUL	Permafrost	ACDPS	Zuse	EERAC	FRDR	Jisc RDSS
Geographic Region	Western Canada	Ontario, Canada	National Canada	State of Berlin	East of England	National Canada	National UK
Domain	Member Academic Research Libraries	Member Academic Research Libraries	Members of CCA	LAM / Research Data in Berlin	Members of EERAC	Research Data Producers in Canada	Research Data UK
Contractor	COPPUL	OCUL	CCA	Zuse	EERAC	Compute Canada*	Jisc
Status	Production	Pilot (Q2 '18)	Production	Production	Completed	Production	Beta (Q3 '18)
Platform vendor	Educloud	OCUL	OVH	Zuse	Amazon	Globus	Amazon

Consortia Shared Services

Service	COPPUL	Permafrost	ACDPS	Zuse	EERAC	FRDR	Jisc RDSS
User Login	open	open	open	closed	open	closed	open or closed
Resources	dedicated	dedicated	dedicated	shared	shared	shared	dedicated or shared
platform	VMWare	OpenStack	OpenStack	kvm	AWS	Globus	ECS
End user support	Artefactual	OCUL	Artefactual	–	Arkivum	–	Jisc
Distributed storage	Educloud	OLRC	Azure	iRODS	Arkivum	Compute Canada	Arkivum/ UK Cloud
Access	AtoM/Local	Local	AtoM	Fedora4	AtoM	frdr.ca	RDSS

Links

ACDPS	Archives Canada Digital Preservation System	EERAC	East of England Regional Archive Council
Arkivum	Arkivum Perpetua	FRDR	Federated Research Data Repository
Artefactual	Artefactual Hosted Services	Globus	Globus Grid FTP
AtoM	Access To Memory	iRODS	Open Source Data Management Software
AWS	Amazon Web Services Platform	OCUL	Ontario Council of Research Libraries
Azure	Microsoft Azure Cloud Services Platform	OLRC	Ontario Library Research Cloud
CCA	Canadian Council of Archives	OpenStack	Open Source Cloud Platform
COPPUL	Council of Prairie and Pacific University Libraries	Preservica CE	Preservica Cloud Edition
ECS	Amazon Elastic Container Services	RDSS	Research Data Shared Service
EduCloud	University of B.C. Private Cloud Service	Zuse	Zuse Institute Berlin

Ecology of Infrastructure

Ecology

the branch of biology that deals with the relations of organisms to one another and to their physical surroundings.

- Oxford English Dictionary

What is Infrastructure?

Common metaphors present it as a substrate: something upon which something else “runs” or “operates”, [e.g.,] railroad tracks upon which rail cars run. Infrastructure in this image is something built and maintained, sinking then into an invisible back-end.

Such a metaphor is neither useful nor accurate.

Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 253-264. DOI=http://dx.doi.org/10.1145/192844.193021

What is Infrastructure?

… We hold that infrastructure is fundamentally and always a relation, never a thing. … This inversion de-emphasizes things or people as the only causes of change, and focuses on infrastructural relations (e.g. between railroads, timetables, and management structures in bureaucracies).

Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 253-264. DOI=http://dx.doi.org/10.1145/192844.193021

What is Infrastructure?

Most respondents liked the system, praising its ease of use and its understanding of the problem domain. On the other hand, most have not signed on;

Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 253-264. DOI=http://dx.doi.org/10.1145/192844.193021

What is Infrastructure?

Despite good user feedback and user participation in the system development, there were unforeseen, complex challenges to usage involving infrastructural and organizational relationships.

Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 253-264. DOI=http://dx.doi.org/10.1145/192844.193021

What is Infrastructure?

We see these problems not in terms of "user resistance" or "system success/failure." Rather, they are organizational and learning challenges . . .

Susan Leigh Star and Karen Ruhleder. 1994. Steps towards an ecology of infrastructure: complex problems in design and access for large-scale collaborative systems. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (CSCW '94). ACM, New York, NY, USA, 253-264. DOI=http://dx.doi.org/10.1145/192844.193021

Ecology of Infrastructure

- Developing shared services was hard in 1994 – still hard now
- Tackle complex problems with multidisciplinary teams
- Requires communication and collaboration across domains
- Requires time to grow – system sophistication, user capacity

Ecology of Digital Preservation

Software or Data Carpentry is a useful paradigm, but . . .

- Frames problem as completing a ‘finished product’ that is ‘built to spec’

Try thinking instead in terms of Gardening or Farming

- Start a sustainable process which depends on many factors to succeed, most of which are out of your control
- Success is successive – start now and plan for generations