simpleArchive as a Service

Marius Politze

RWTH Aachen University
IT Center
Content

• Challenge: How to get researchers to archive their data?

• Our solution: make it simple
  – simpleArchive concept
  – Demo

• Scaling simpleArchive as a service

• Conclusion and future challenges
Publications, Data, Metadata – A Research Data Infrastructure

Publications
- Nachweis
- Volltext
- Verweis

Data

Metadata of research data

 visibility

+ +

Archive

PID

Metadata Store

RWTH Publications

simpleArchive as a service
Marius Politze
Kolloquium der Betreiber von Forschungsdateninfrastrukturen in NRW, 15.12.2017
Archiving Until Now

- https://doc.itc.rwth-aachen.de/display/ARC/Archiv+Knoten+anlegen
- https://doc.itc.rwth-aachen.de/display/ARC/TSM+Installation
- https://doc.itc.rwth-aachen.de/display/ARC/TSM+Konfiguration+-+Archiv
- https://doc.itc.rwth-aachen.de/display/ARC/Benutzung+des+TSM-Clients
### Requirements

- **Allow researchers to archive “small” files**
  - Up to 2GB
  - Make it a free service so researchers will use it
  - Reduce costs by storing on tape

- **Reuse existing concepts and applications**
  - Allow use in federated context
  - Reduce development and maintenance costs by using available systems

- **Make sharing of archived data as easy as archiving**
  - Archived data is not necessarily open access
  - Let researchers restore their data
  - … and let them share it using a simple URL

- **Make archived data globally identifiable using PIDs**
  - So researchers can reference it elsewhere
  - … and can retrieve it using the PID
Archiving with simpleArchive

![Screenshot of the simpleArchive interface]

simpleArchive - Einfach und schnell Dateien archivieren

Die Anwendung ermöglicht es, Dateien zu archivieren und aus dem Archiv wiederherzustellen.

- Datei archivieren
- Datei wiederherstellen

Eine Datei ins Archiv hochladen

Wiederherstellung einer bereits archvierten Datei

Bei Fragen oder Anregungen wenden Sie sich gerne an service@itc.rwth-aachen.de.
Archive and Restore Process (simplified)

**User**
- upload file
- notify user

**Temp. File System**
- save file
- create PID

**ePIC**
- create PID

**Timestamp**
- sign file hash
- schedule archival
- archive file

**Tape Archive**
- schedule restore
- retrieve file

**Notify User**
- request file
- notify user

- create temporary download
simpleArchive is an implementation of a process not an application!
Concept: Software Layers

- Base applications and services:
  - Sciebo mit FDM Erweiterungen
  - Zenodo/Invenio
  - Datenmanagementpläne
  - Metadatentool
  - PID
  - PID

- Common Userinterfaces:

- Common Processes:

- IdM / Roles / DFN-AAI:
  - Virtualized Compute
  - Object Store
  - ISP
  - Rosetta

- IdM / Roles / DFN-AAI:
  - Private Domäne
  - Gruppen-domäne
  - Dauerhafte Domäne
  - Zugriff & Nachnutzung
Loadbalancer and internet connection to DFN Network
- DNS Loadbalancing
- Redundant sites in Aachen (SW23 and WW10)
- Redundant connection to DFN Network

User Interface: app.rwth-aachen.de
- Shared Hosts with process layer
- Accesses process layer via load balancers

Processes: moped.ecampus.rwth-aachen.de
- 4 VMs at Redundant in sites Aachen (SW23 and WW10)
- Each site retains capacity to keep services available in case of site failure
- Homogeneous access to base applications and services
- Automated deployment

Base applications and services
- Base on specific OLAs with the service providers
- Partially redundant, cold standby or disaster recovery
- Failures in these systems impact only dependent processes
Scaling Out: Vision 2018

Providing FDM Processes and Infrastructure as a service

• **Pro**
  – Simple for customers and providers
  – Only single instance reduces maintenance costs
  – Reuses already available federated infrastructures like DFN-AAI

• **Con**
  – Failure in the instance impacts all customers
  – Does not scale for data or compute intensive services
  – Researchers and service providers often want to keep services local
Scaling Out: Vision 2018

Scaling by adding new sites

- **Pro**
  - Mirroring infrastructure components increases redundancy
  - Local services remain for local users and
  - Services can be used cross-site

- **Con**
  - Maintaining multiple infrastructures becomes expensive
  - Instead of core scientific processes sites may degenerate to support only local services
Scaling Out: Vision 2018

Scaling by adding base applications and services from other sites

- **Pro**
  - Compute and data capacity provided locally
  - Easy cross-site reuse of services
  - Using available federative infrastructures
  - Standardized processes allow interoperability

- **Con**
  - Failure in process layer impacts all users
  - OLAs required to control users and processes
Conclusion & Future Challenges

• simpleArchive is available to all researchers at RWTH Aachen since Q3 2016

• Implementation of process reuses existing systems and APIs

• Focusing on the process rather the technology reduces vendor-lockin

• Process needs to be backed by local policies
  – How long is the data actually stored?
  – Who can restore the data?
  – Can archives be transferred?
  – Can archives be deleted?

• Combine scaling methods to build a process oriented cloud-like ecosystem
Thank you for your attention

Vielen Dank für Ihre Aufmerksamkeit