



HAL
open science

Experiment Data Management

Lucas Nussbaum

► **To cite this version:**

Lucas Nussbaum. Experiment Data Management. GEFI 2018 - Global Experimentation for Future Internet, Oct 2018, Tokyo, Japan. hal-01944472

HAL Id: hal-01944472

<https://hal.inria.fr/hal-01944472>

Submitted on 4 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Experiment data management

Lucas Nussbaum

GEFI workshop 2018



Motivations

- ▶ **Do the right thing:**
Proper sharing of research artifacts is a requirement for reproducibility
 - ◆ Details of SuT
 - ◆ Experiment orchestration code
 - ◆ Input & output data
- ▶ **Hype** around *Open Science*
 - ◆ Publications of course, but also data
- ▶ **Policies**
 - ◆ Requirements for Data Management Plans
 - ◆ *Generated data volume* as a metric for the usefulness of RIs

Requirements

- ▶ Storing data *during* experiments \leadsto usually on nodes
- ▶ Storing data *between* experiments
- ▶ Archiving data after experiments
 - ◆ With a stable reference (DOI)
 - ◆ On the very long-term
- ▶ Data Management Plan

Status

Data Management Plan:

- ▶ *Testbed X does not produce data itself (not a telescope). User experiments do, they should be the ones with a DMP.*
 - ◆ Hard sell in the Research Infrastructures community

Status

Data Management Plan:

- ▶ *Testbed X does not produce data itself (not a telescope). User experiments do, they should be the ones with a DMP.*
 - ◆ Hard sell in the Research Infrastructures community

Archiving:

- ▶ De-facto standard: GitHub (and not very well done – commit hash?)
- ▶ What we should probably explore: data repositories
 - ◆ Public instances: Zenodo, Figshare, Driad, KNB, ICPSR
Limitations: size, no statistics
 - ◆ Self-hosted: Dataverse, CKAN, Fedora Commons (usually hosted by universities – useful to have instances for testbeds or federations?)

Status

Data Management Plan:

- ▶ *Testbed X does not produce data itself (not a telescope). User experiments do, they should be the ones with a DMP.*
 - ◆ Hard sell in the Research Infrastructures community

Archiving:

- ▶ De-facto standard: GitHub (and not very well done – commit hash?)
- ▶ What we should probably explore: data repositories
 - ◆ Public instances: Zenodo, Figshare, Driad, KNB, ICPSR
Limitations: size, no statistics
 - ◆ Self-hosted: Dataverse, CKAN, Fedora Commons (usually hosted by universities – useful to have instances for testbeds or federations?)

Storage:

- ▶ ChameleonCloud: Swift-based object store
- ▶ Emulab: per-project NFS dir (no quotas on Virtual Wall, 100 GB on CloudLab)
- ▶ Grid'5000:
 - ◆ NFS directory per user (now with more security)
 - ◆ **Disk reservation**

Disk reservation on Grid'5000¹

Target experiments:

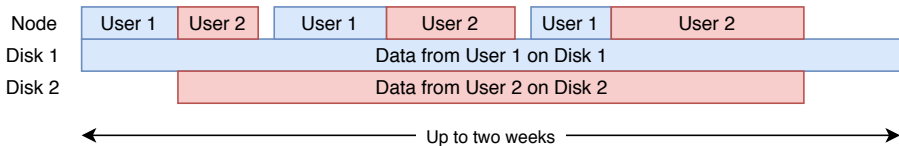
- ▶ Evaluation/use of Big Data solutions

Problem:

- ▶ Moving data to/from nodes is time-consuming
- ▶ Amplified by our short reservation policy (one night / one week-end)

Solution:

- ▶ Make it possible to leave data on nodes (using additional disks)



- ▶ Implemented by enabling/disabled disks in the RAID controller based on who reserved the node
 - ◆ Disks are not visible by other users

¹Joint work with Florent Didier and Pierre Neyron