Zenodo: A Research Data Repository for All

ICSTI TACC Workshop – 2017/04/04
Scientific Data Repositories: Use Cases, Innovation, and Best Practices
Data ...Mountain...Lake...Cloud...

8 GB/s

150 PB
150,000,000 GB

400 TB
Open Data as a Service

Open Access Mandate
Open Data Pilot

 zenodo

OpenAIRE
Findable & Interoperable

Findable & Interoperable

Grant DB: FP7, H2020
…WT (UK), FCT (Portugal), ARC & NHMRC (Australia)
Communities

Projects, Subjects, Institutes, Nations, Conferences, …
Open Scholarship: Share ≠ Publish ≠ Preserve
- **Upload Big Data**
  - Stream binary content, no encoding/decoding of file at either end
  - Received and streamed directly to storage cluster
  - Tuned js client, web server, application server, application, storage cluster, ...

- **Download Big Data**
  - EOS high-capacity low-latency disk-based storage system
  - Example CODP: XRootD protocol to access parts of data "on demand"
Accessible: Permissions

- Closed access
  Publication moment is too late!
  Capture before the knowledge evaporates
  While data is live

- Restricted access
  Share unpublished data with reviewers
  Share sensitive data with controlled list

- Embargoed access

- Open access

Open Data Repository ≠ Simple
  Logins & Roles
  Fine grained ACLs
  Sophisticated security model
  Life-cycles and workflows
Keeping Data Accessible

- **Media Verification**
  - Hot / Cold Data
  - 10% of production drive capacity for 2.6 years

- **Media Migration**
  - Drive and Media obsolescence
  - 50% of current drive capacity for 2 years

O($10^{-16}$) bit error rate
Keeping Data Accessible

Total Tape Reading, (TB)

- Verification
- Data Recording / Analysis
- Repack

Years: 2009 to 2015
• Building system to **instantiate** preserved analysis on the cloud

• Supporting **multiple**…
  – computing clouds
  – running environments
  – resource orchestration
  – workflow engines
  – shared storage systems

Preservation: Driven by Science

- Bit Preservation: ISO 16363 standard
- Software preservation
- Analysis Preservation
- REusable ANAlyses
- Zenodo > Data Seal of Approval
Concluding Remarks

• Data Management best done by the data owner
  – Make data management services available
  – Empower the users!
• Don’t assume Researcher’s behavioral patterns
  – Decouple the layers
  – Offer integratable building blocks, not an integrated solution
• Housekeeping (media verification and migration) takes non-negligible resources
• Preservation means combining expertise
  – Storage managers, SW engineers, Librarians and Researchers
Thank You! Questions?

Tim.Smith@cern.ch
0000-0002-1567-7116
http://cern.ch/tim.smith

http://zenodo.org
http://www.openaire.eu
http://invenio-software.org
http://github.com/reana
Lars.Nielsen@cern.ch