EST
Data Rates & Volumes

P. Caligari (KIS), N. Bello González (KIS), S. Berdyugina (KIS), P. Kehusmaa (KIS)

EST-DC Preliminary Requirements

1. **Data Policy** (decision needed before starting designing!)

- Data **needs a License!** (e.g.: https://creativecommons.org)
  - CC0 1.0 (public domain)
  - CC BY 4.0 (credit, share & adapt, commercial usage allowed)
  - The more restrictive a license the more problematic it becomes to combine with data from other sources

use for metadata

use for freely accessible data

- Ownership?
  - EST / Consortium?
- Copyright?
- Embargoes?
  - How long by default?
  - Longer if PhD involved?
  - What about 3rd party campaigns?
  - What about technical campaigns, ad-hoc campaigns, exceptions from the default?

SDC: 1 year

SDC: 2 years

- We should form a workgroup formulating drafts, asap!

Terminology

- Define **data levels**

SDC

- L0: raw data & files needed for calibration
- L1: calibrated data (science ready)
- L2: reduced data, data products (inversions, speckled data, etc.)
- L3: higher level data products (e.g. Statistical analyses)

IVOA

- L0: raw instrumental data requiring instrument-specific tools
- L1: instrumental data processable with standard tools
- L2: calibrated, science-ready data without instrument signature
- L3: enhanced data products (e.g., mosaics)

- Agree on **data format** (ideally all levels)
- Agree on **metadata standard** & define **minimal header** (instrument **in**dependent)
- Agree on **file naming**
- Agree on **terminology** (instrument dependent?)
  - What is reduction?
  - What is calibration?

Data Rates & implications

- **Claudia: ~14 PB/d (LL, 4h observation)**
  - => 1 mil€/d[1)] just to store on disk (assuming: 64 TB disks, 2 copies, 2/3 disk redundancy), not considering:
    - Costs for computational hardware
    - Cooling/electricity costs
  - Peak days with observations > 4h drive that costs up at the summit!
- **Long term storage (on the continent)**
  - "keep raw data for some years" => 1.3 EB/y (LL, 100 observation days)
    - => 100 mil€/y (not considering storing calibrated & reduced data and higher level data products)
    - Some might be kept on tape (factor 2 cheaper than disks; not near-line, though)
    - Guess cloud will still be more expensive than disks (traffic costs)
- **Data transport to the continent requires:**
  - 14 PB/d within 24h: 1.3 Tb/s Line!
  - 1.3 EB/y within 365d: 400 Gb/sec

: [1)]https://wolke7.leibniz-kis.de/s/7TXAdJtfnFzBZ88

Costs & Concrete Design

- **Per site:**
  - Estimate **building costs**; needs a reliable model of:
    - Storage requirements
    - Computing Requirements (difficult!)
  - Based on these requirements, estimate **operational costs**, including:
    - Long term storage
    - Producing standard higher-level data products (if wanted?)
    - Staffing
    - Hardware renewal
    - Energy costs
  - When needed?
- **To make a long story short: I am not worried about (technical) feasibility, I'm worried about:**
  - Costs
  - Implications for infrastructure (buildings, cooling, etc.)
- **Time to talk about trade-offs?!**

It’s time to form a DC working group!

Who is interested?

Data Management & Life Cycle

Excursion to Rucio[1)]

- One data lake consisting of
  - Multiple geographically distributed sites
  - Multiple Rucio Storage Elements (RSEs) per site
- Write-Once-Read-Many storage (WORM)
  - Files cannot be modified (versioning in the filename needed), only deleted (automatically I lifetime expired and nobody claims an interest)
  - File names cannot be reused!
  - Data Identifiers (DIDs): *scope:name*
  - Flat Namespace, but grouping is possible (**virtual** folders): DIDs -> *datasets* & *datasets/containers* -> *containers*
- Associate human-readable Tags to sites & RSEs
- Data Management (Rucio) by tag-aware *rules*, e.g.:
  - Keep 2 copies of *scope:name* at *location=CALB* for *3 months*
  - *Keep 1 copy of* *:* at *location=PDC&type=tape* (no lifetime = forever)
  - Default Rules per data-source (identified e.g. by *scope*) & users rules possible
  - Nice trick
    - associate DOI to *container*
    - *Rule:* keep *DIDs* in *container* for *10 years*

[1)] https://doi.org/10.1007/s41781-019-0026-3

Thanks.