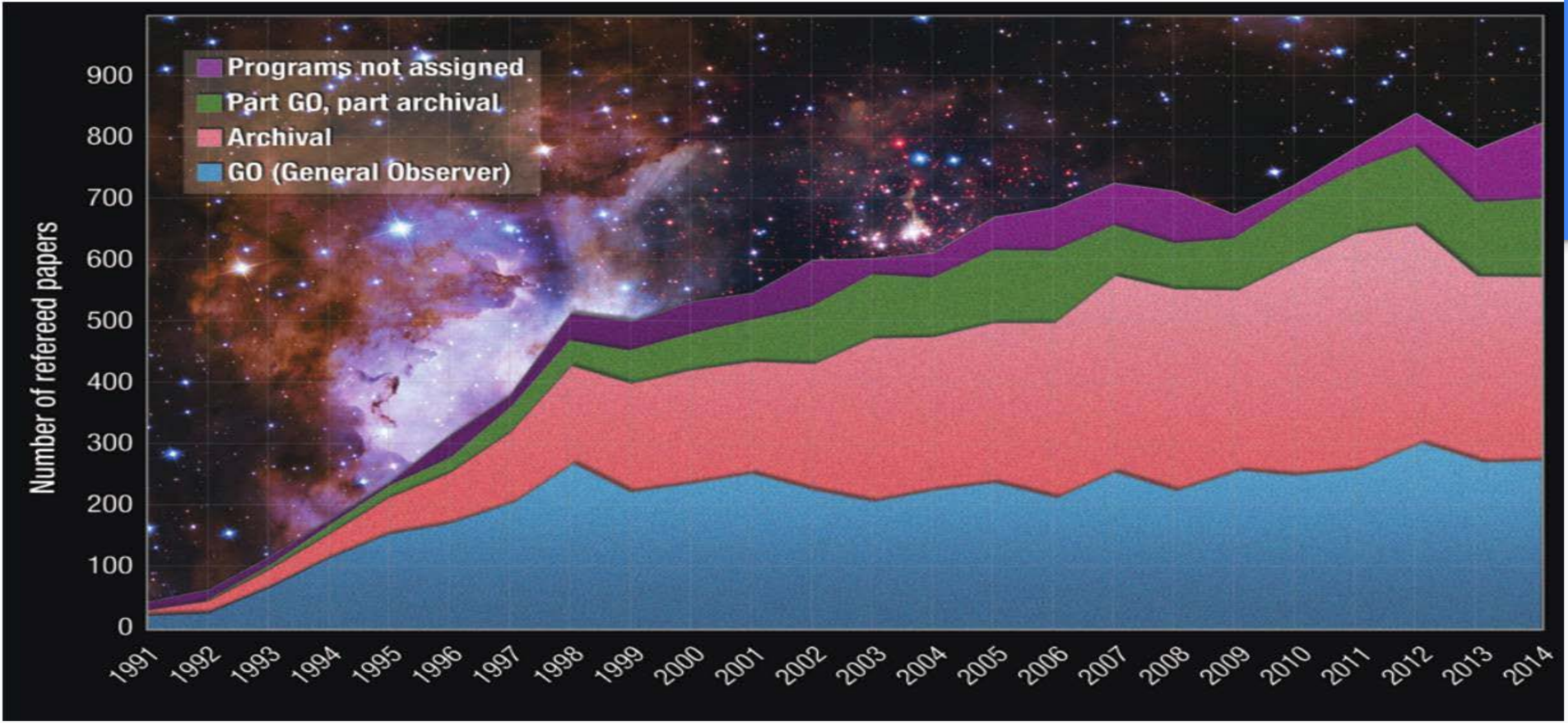
LBT Archive:
…not only save the bit

R. Smareglia, C. Knapic and the IA2 team

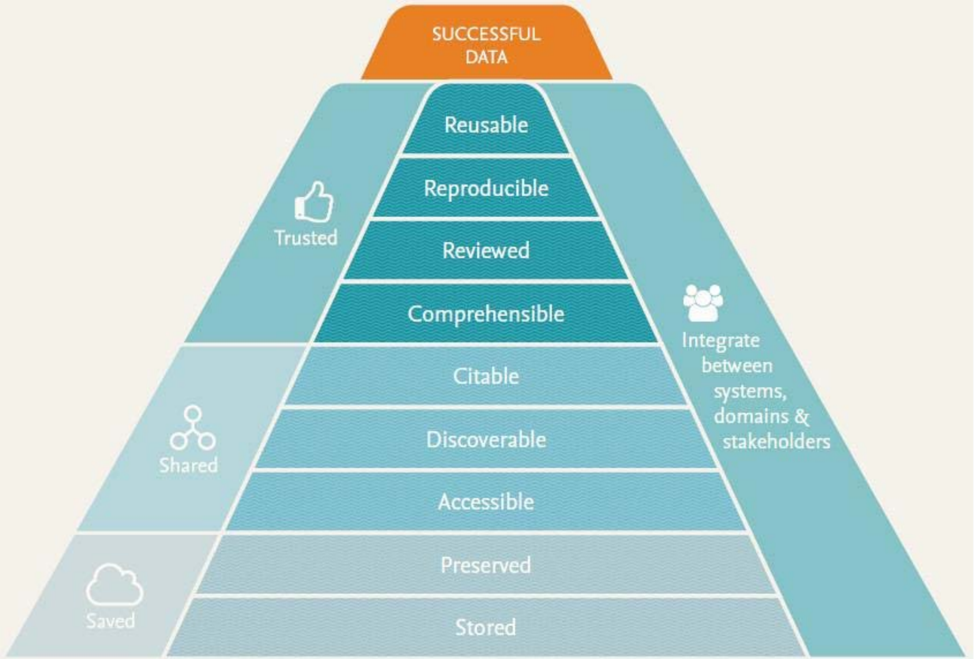


HST Newsletter: “*At the present time, approximately **half of the refereed publications** based on Hubble observations are derived purely **from archival data**, and, every year, this number is slightly higher than the number of publications based on new observations.*
.... the Hubble Archive has become a goldmine for the astronomical community....”

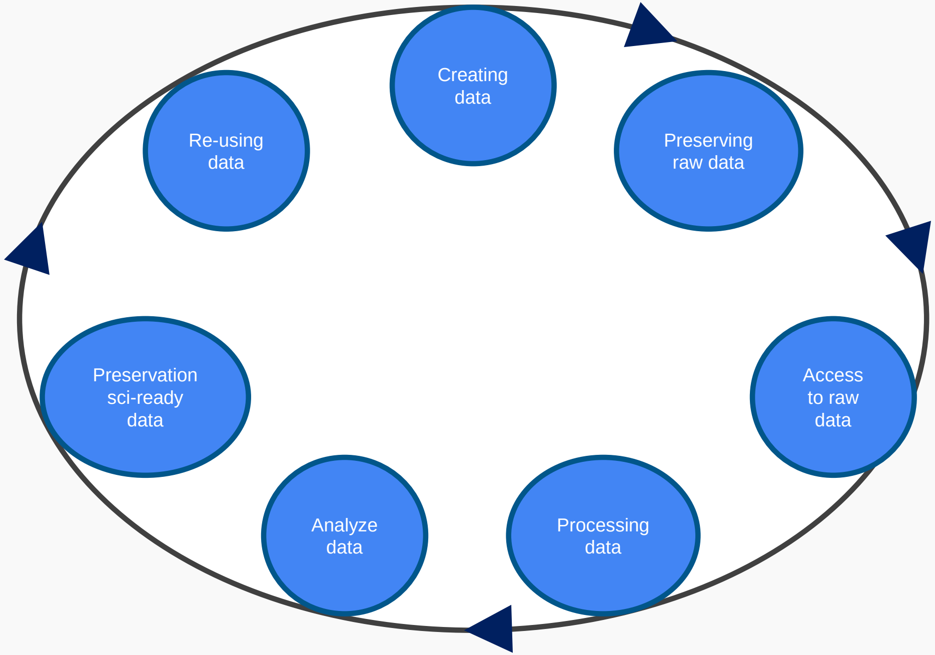
Data as virtuose circle

Two questions:

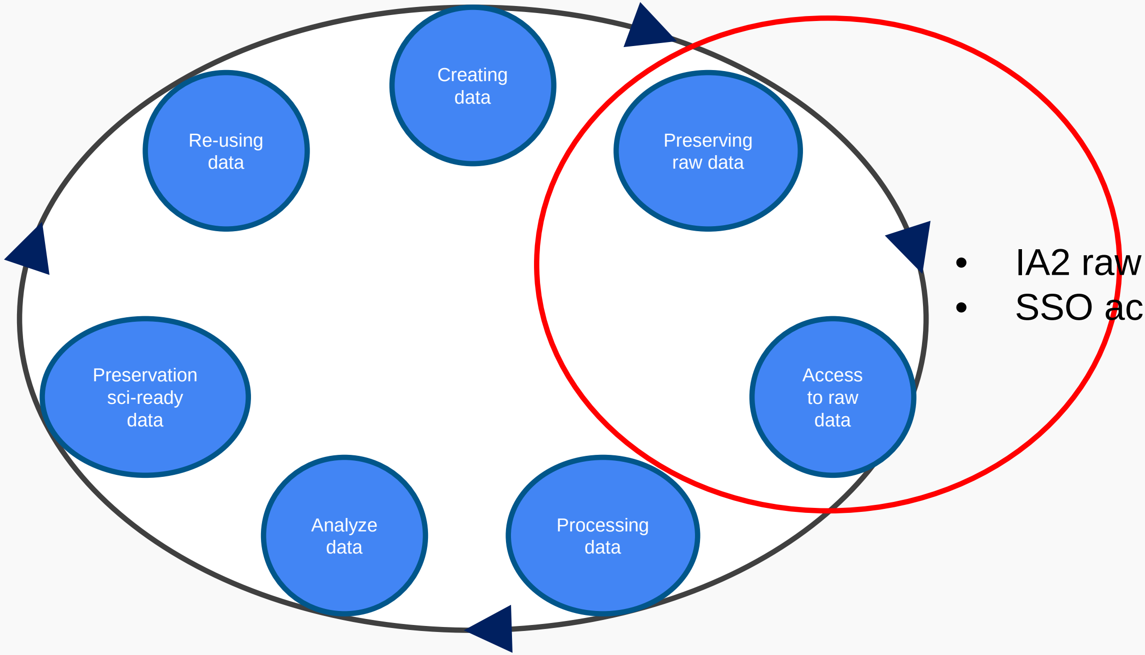
- How to Maximize scientific output of LBT
- Publish more from **their** LBTO data



Research data (Virtuose) lifecycle

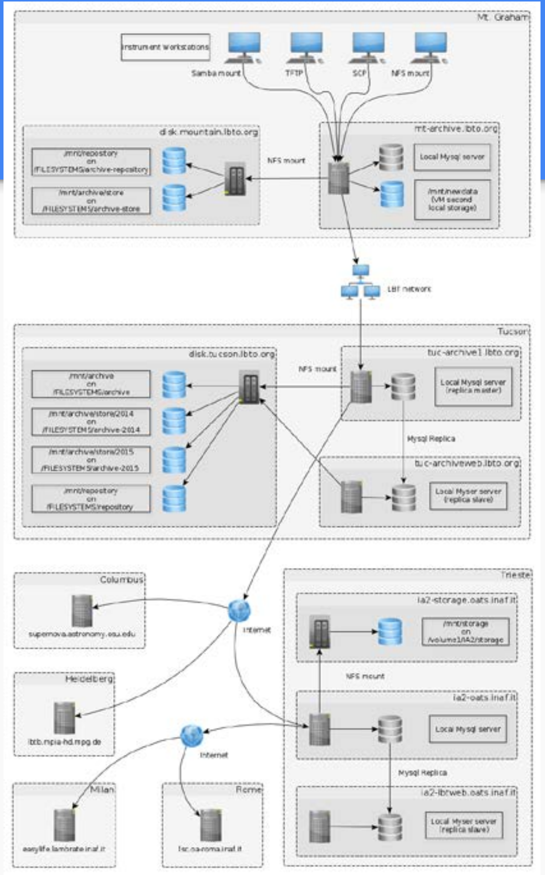


Research data lifecycle



- IA2 raw data archive
- SSO access to data

Archive raw data: status



Mt. Graham
Instrument Workstations
Samba mount
TFTP
SCP
NFS mount
disk.mountain.lbto.org
/mnt/repository on /FILESYSTEMS/archive-repository
/mnt/archive/store on /FILESYSTEMS/archive-store
NFS mount
mt-archive.lbto.org
Local Mysql server
/mnt/newdata (VM second local storage)
LBT network
Tucson
disk.tucson.lbto.org
/mnt/archive on /FILESYSTEMS/archive
/mnt/archive/store/2014 on /FILESYSTEMS/archive-2014
/mnt/archive/store/2015 on /FILESYSTEMS/archive-2015
/mnt/repository on /FILESYSTEMS/repository
NFS mount
tuc-archive1.lbto.org
Local Mysql server (replica master)
Mysql Replica
tuc-archiveweb.lbto.org
Local Myser server (replica slave)
Columbus
supernova.astronomy.osu.edu
Internet
Heidelberg
lbtb.mpia-hd.mpg.de
Internet
Milan
easylife.lambrate.inaf.it
Rome
lsc.oa-roma.inaf.it
Trieste
ia2-storage.oats.inaf.it
/mnt/storage on /volume1/IA2/storage
NFS mount
ia2-oats.inaf.it
Local Mysql server
Mysql Replica
ia2-lbtweb.oats.inaf.it
Local Myser server (replica slave)

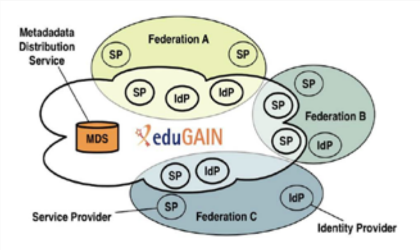
- IA2 data center Manage LBT corporation data since… beginning

- Data: about 650K images ~12 TB

- Data Access from Tucson + Trieste
  - From 2010:
    - Access: 134 K (41 + 93)
    - Download: 26.5 K (9.5 + 17)
  - Last year:
    - Access: 33 K (6 + 27)
    - Download: 9.3 K (1.2 + 8.1)

Archive status: evolution to SSO

Authentication



IA2 could very soon offer a SSO integrated mechanism to authenticate.

It is in beta version at the IA2 data center.

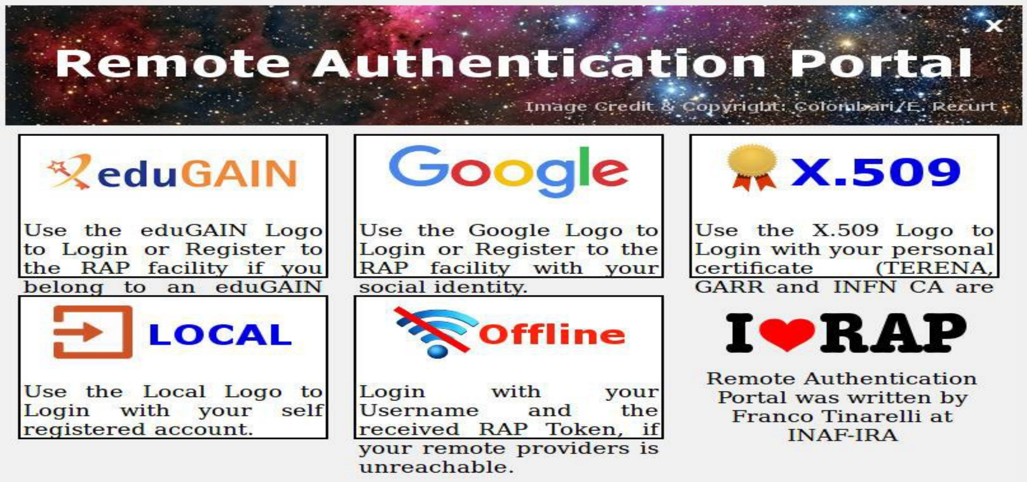
It will support the most diffuse and reliable authentication systems and will manage the user authorizations via a Grouping Management system.

Courtesy of F. Tinarelli

Authorization:

Internet2 application based / VO compliant based

Grouper™



Remote Authentication Portal

Image Credit & Copyright: Colombari/E. Recurt

eduGAIN

Use the eduGAIN Logo to Login or Register to the RAP facility if you belong to an eduGAIN

Google

Use the Google Logo to Login or Register to the RAP facility with your social identity.

X.509

Use the X.509 Logo to Login with your personal certificate (TERENA, GARR and INFN CA are

LOCAL

Use the Local Logo to Login with your self registered account.

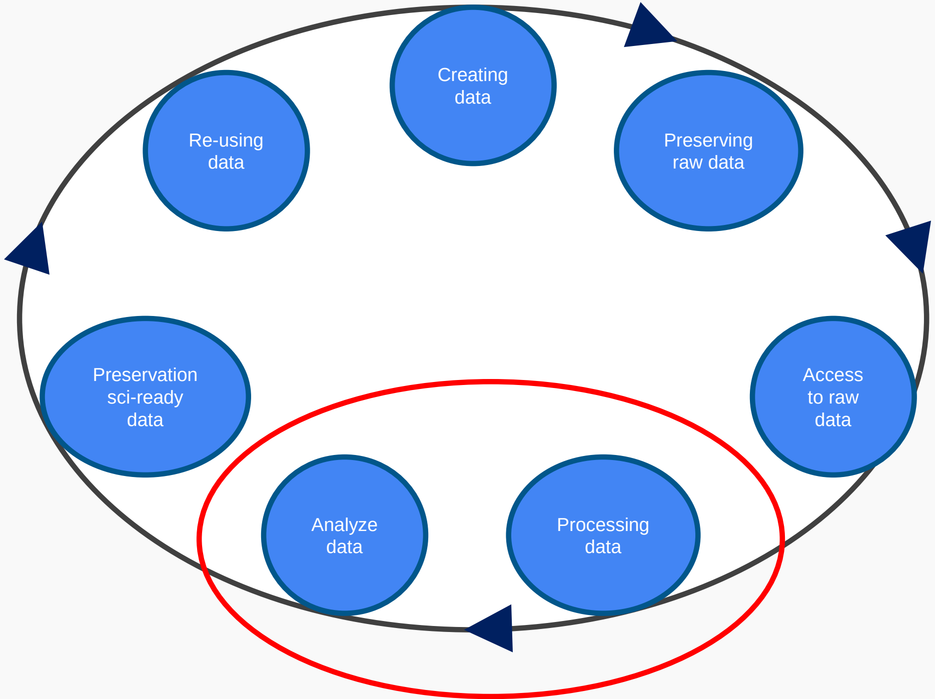
Offline

Login with your Username and the received RAP Token, if your remote providers is unreachable.

I ❤ RAP

Remote Authentication Portal was written by Franco Tinarelli at INAF-IRA

Research data lifecycle



- Cloud middleware
- Cloud archive
- ~~Analysis sw~~

Data processing: ex. of INDIGO project

INDIGO and the EU Open Science Cloud

INDIGO-DataCloud develops an **Open Source data and computing platform** provisioned over private, public or hybrid e-infrastructures.

INDIGO - DataCloud
Better Software for Better Science

By **filling gaps** of current Cloud technologies [see next slides], INDIGO- DataCloud helps scientists, software developers, resource providers and e-infrastructures to **efficiently exploit computing, data and network technologies**:

Better Software for Better Science.

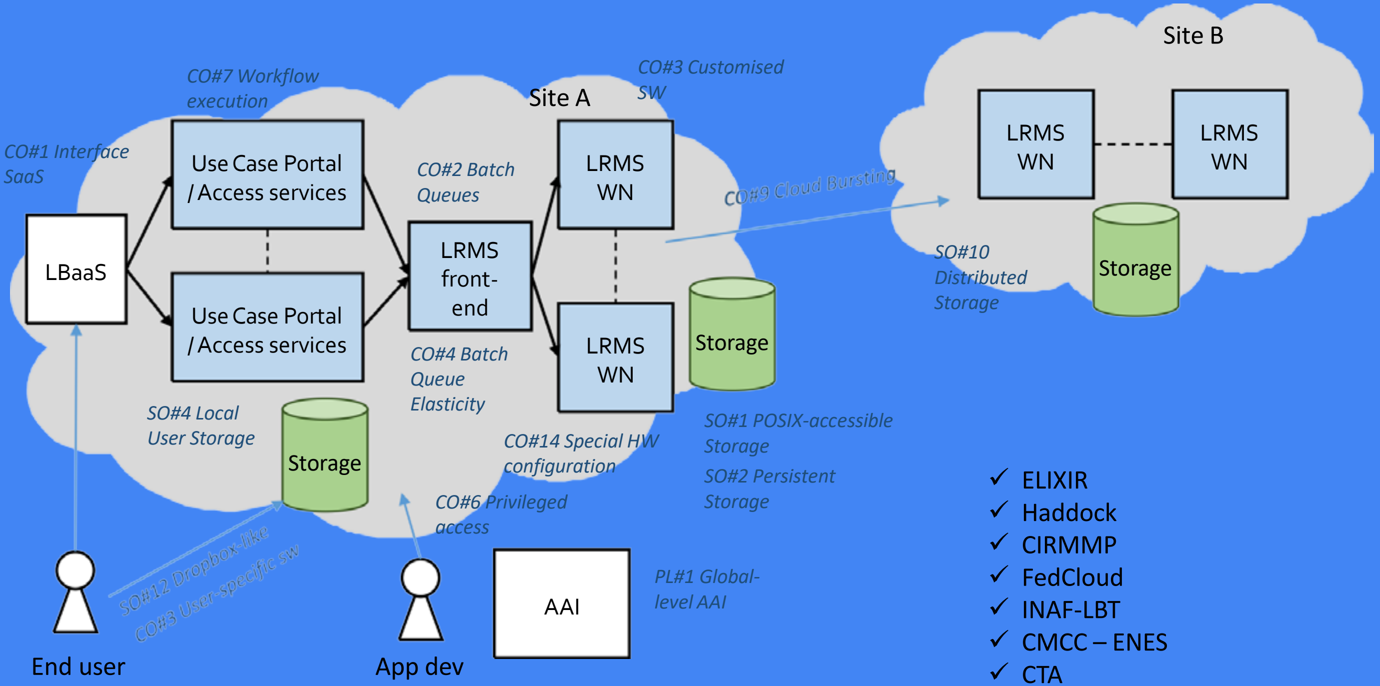
INDIGO - DataCloud
Better Software for Better Science

INDIGO project: User as first

- **Requirements come from research communities**
    - “The proposal is oriented to support the use of different e- infrastructures by a wide-range of scientific communities, and aims to address a wide range of challenging requirements posed by leading-edge research activities conducted by those communities.” (INDIGO DoW)
- **We gathered use cases** from 11 scientific communities.
    - LifeWatch, EuroBioImaging, INSTRUCT, **LBT**, CTA, WeNMR, ENES, eCulture Science Gateway, ELIXIR, EMSO, DARIAH.
- Starting from about 100 distinct requirements we derived a much shorter list, grouped **into 3 categories: Computational requirements, Storage requirements, Requirements on infrastructures**. Each requirement has an associated ranking (mandatory / convenient / optional).
- See **https://www.indigo- datacloud.eu/pages/components/deliverables.html**

Computational Portal as a Service

INDIGO - DataCloud
Better Software for Better Science



- ✓ ELIXIR
- ✓ Haddock
- ✓ CIRMMP
- ✓ FedCloud
- ✓ INAF-LBT
- ✓ CMCC – ENES
- ✓ CTA
- ✓ ALGAE – BLOSSOM

INDIGO - DataCloud
Better Software for Better Science

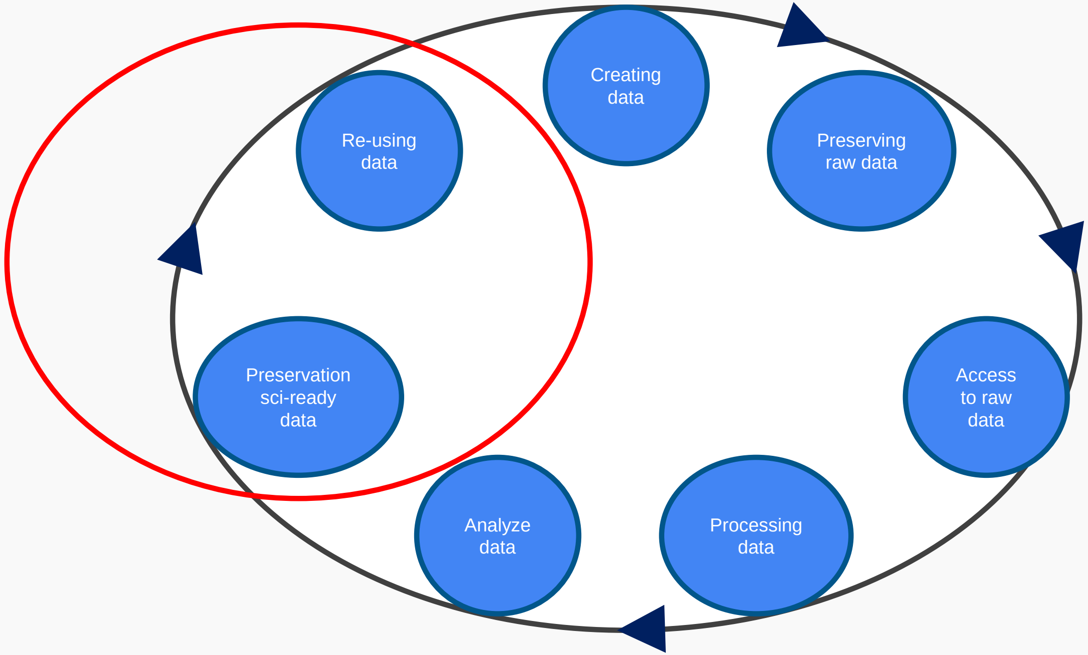
INDIGO lesson learned:

INDIGO, as any Cloud system, is a system made for computing system (as SaaS) but is not a repository or a archive system.

Can be useful for “experimental” data shared between groups (short or large) but is not able to manage data as necessary in an Observatory and for long term preservation and curation.

Can be useful as environment (like many other) where running pipeline (many tools about data sharing, workflow management System, multi versioning, …)

Research data lifecycle



- Virtual Observotry
- Open data
- Open science
- Interoperability
- Intellectual Propriety

Interoperability→ Virtual Observatory →Open Access / Open Science

Open Access and Open Science is one of the ***MUST*** *of the*

EU/H2020 funding project policy

- *The* ***European Open Science Cloud*** *(EOSC) pilot project, in which INAF is involved, will support the first phase in the development as described in the EC Communication on European Cloud Initiatives [2016].*

– It will establish the governance framework for the EOSC and contribute to the development of European open science policy and best practice;

–It will develop a number of pilots that integrate services and infrastructures to demonstrate interoperability in a number of scientific domains; and
It will engage with a broad range of stakeholders, crossing borders and communities, to build the trust and skills required for adoption of an open approach to scientific research

Intellectual Propriety

First of all: ONE data policy (only INAF have a data policy since 2007)

Save Intellectual propriety for:

- Software
- Sci-ready data
- Data raw ?? (based on data policy)

An organized system can save IP using DOI

Studies have found that papers with publicly available datasets receive a higher number of citations than similar studies without available data. …

https://peerj.com/articles/175/

About DOI

DOI (Digital Object Identifier):

DOI is a character string used to uniquely identify an object such as an electronic document. The *DOI* for a document is ***permanent***, whereas its location and other meta-data may change. Referring to an on-line document by its DOI provides more stable linking than simply referring to it by its URL, because if its URL changes, the publisher need only update the meta-data for the DOI to link to the new URL.

- DOI for software (make your Code citable)
- DOI for sci-ready data

CV note .. 2013: measures to enhance science productivity

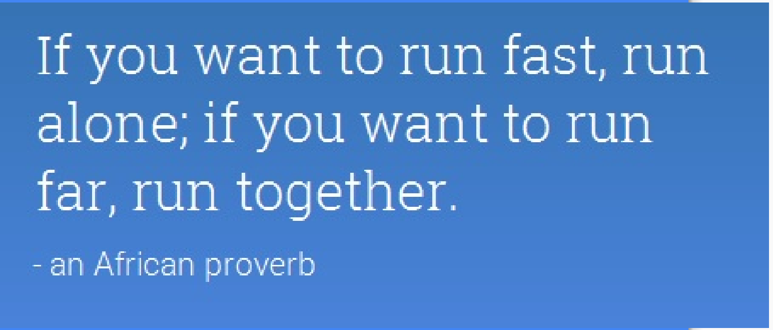
Pipelines already exist for the facility instruments used in seeing limited mode. It is not the intent of LBTO to reinvent them. Instead, we will work with the teams which have developed them in the partnership to bring them to LBT and make them simple and robust enough for an observatory environment. LBTO will maintain them in collaboration with the users' community.

The general purpose of these pipelines would be to:

- remove the instrumental signature in imaging mode (dark, bias, flat-fielding, sky subtraction, possibly PSF reconstruction for AO imaging, ..) with basic astrometry and photometric calibration, and

- produce a calibrated spectrum per slit in spectroscopic mode and a basic standard wavelength and flux table per object on a VO-compatible format, to be subsequently used by applications like VOSpec.

Conclusion



If you want to run fast, run alone; if you want to run far, run together.

- an African proverb

➔ Archive and analysis system <u>**ARE INSTRUMENTS**</u>
(at the moment LBT raw data archive is not yet present in the LBTO web page)

➔ Data Policy is needed to increase scientific return

➔ Open data is an opportunity not a “theft”

➔ All tools (archive, cloud computing, SSO, …) are a reality,
need only to build a single integrate LBT system.