INAF Italian Astronomical Archives facility

17/06/2019 - Cristina Knapic
on behalf of IA2 team

IA2 *Italian Center for Astronomical Archives*
Centro Italiano Archivi Astronomici

INAF ISTITUTO NAZIONALE DI ASTROFISICA

Home | Services | Projects | Software | Additional Info | IA2 Group

ABOUT US

collapse [-]

IA2 (Italian center for Astronomical Archive) is an Italian Astrophysical research e-infrastructure project that aims at co-ordinating different national initiatives to improve the quality of astrophysical data services. It aims at co-ordinating these developments and facilitating access to this data for research purposes. The IA2 is supported by INAF since 2005. IA2's main goals consist in data archiving systems and safety, including data hosting and data curation and preservation, data and metadata distribution over geographical sites, access services including publication within the VO scenario. IA2 provides also services and tool to the community, like data sharing (owncloud), project management (redmine), software collaboration (git-lab) and has available a workflow manager (Yabi) for computational needs.

contact us

MAIN ACTIVITIES

collapse [-]

TELESCOPE ARCHIVES & SIMULATIONS

MAIN: TNG, LBT, Asiago, SVAS, RADIO

MAIN: Exoclimates, INTRIGOSS, LBT OLD

HOSTED: Byurakan, INES, BaSTI, IBIS-A, other archives @ INAF

OTHER SERVICES

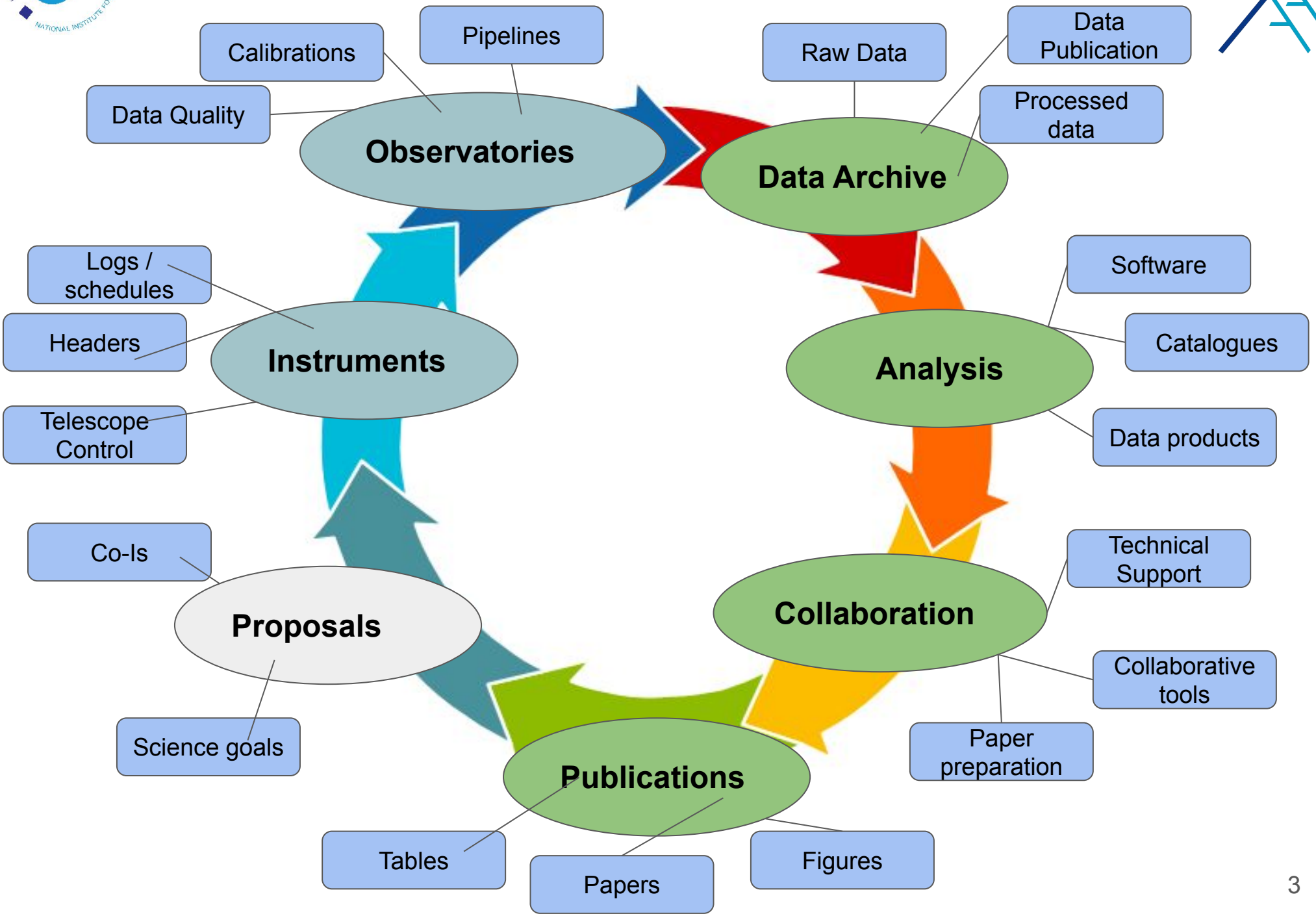
ownCloud, redmine, VO initiatives, GitLab, yabi

PROJECTS

expand [+]

https://ia2.inaf.it/

Astronomical Data Life Cycle

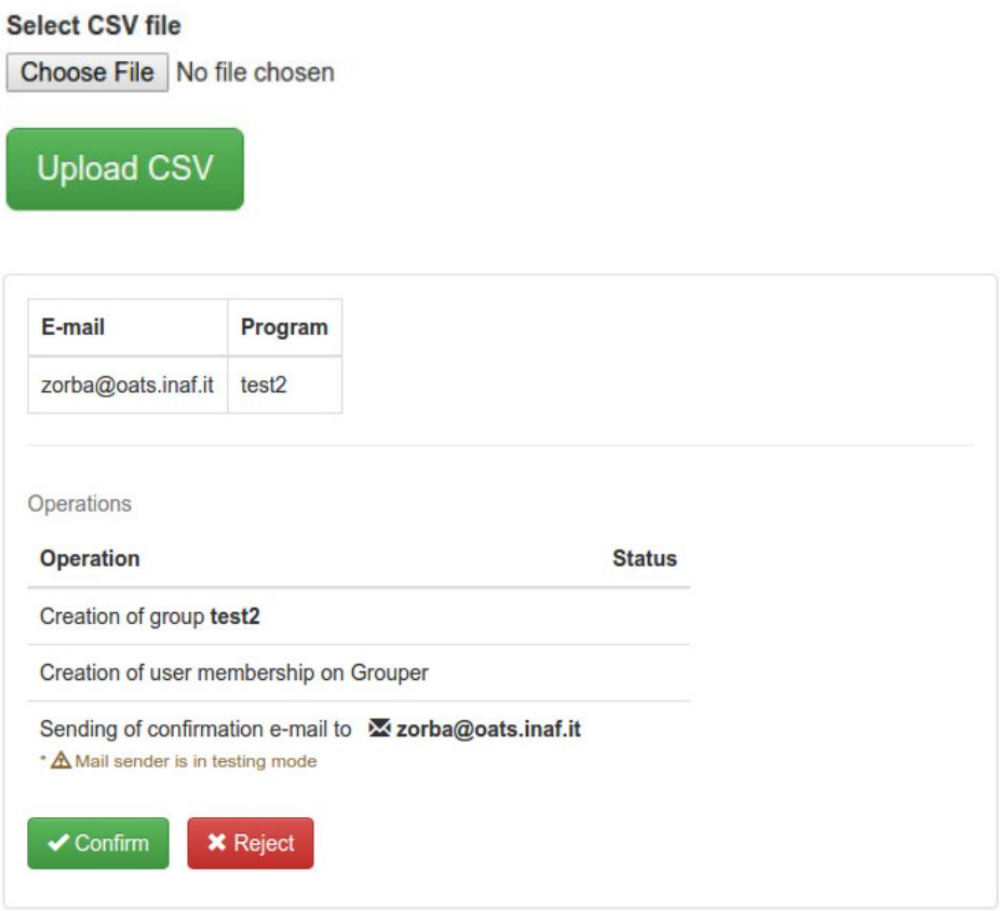


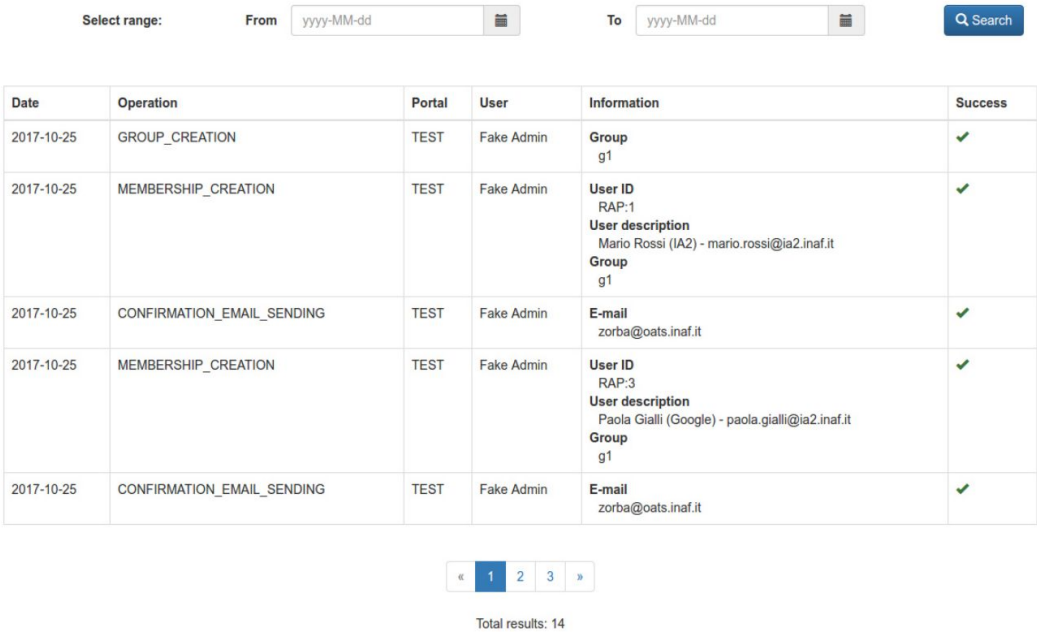
IA2 proposal preparation support

Generation of Observing programs links to registered/known users via administration tool:

CSV format

Program name	User name and surname	E-mail	Write into Postgres (1=true, 0=false)	User address	Schedule time	Priority	Call	Support e-mail

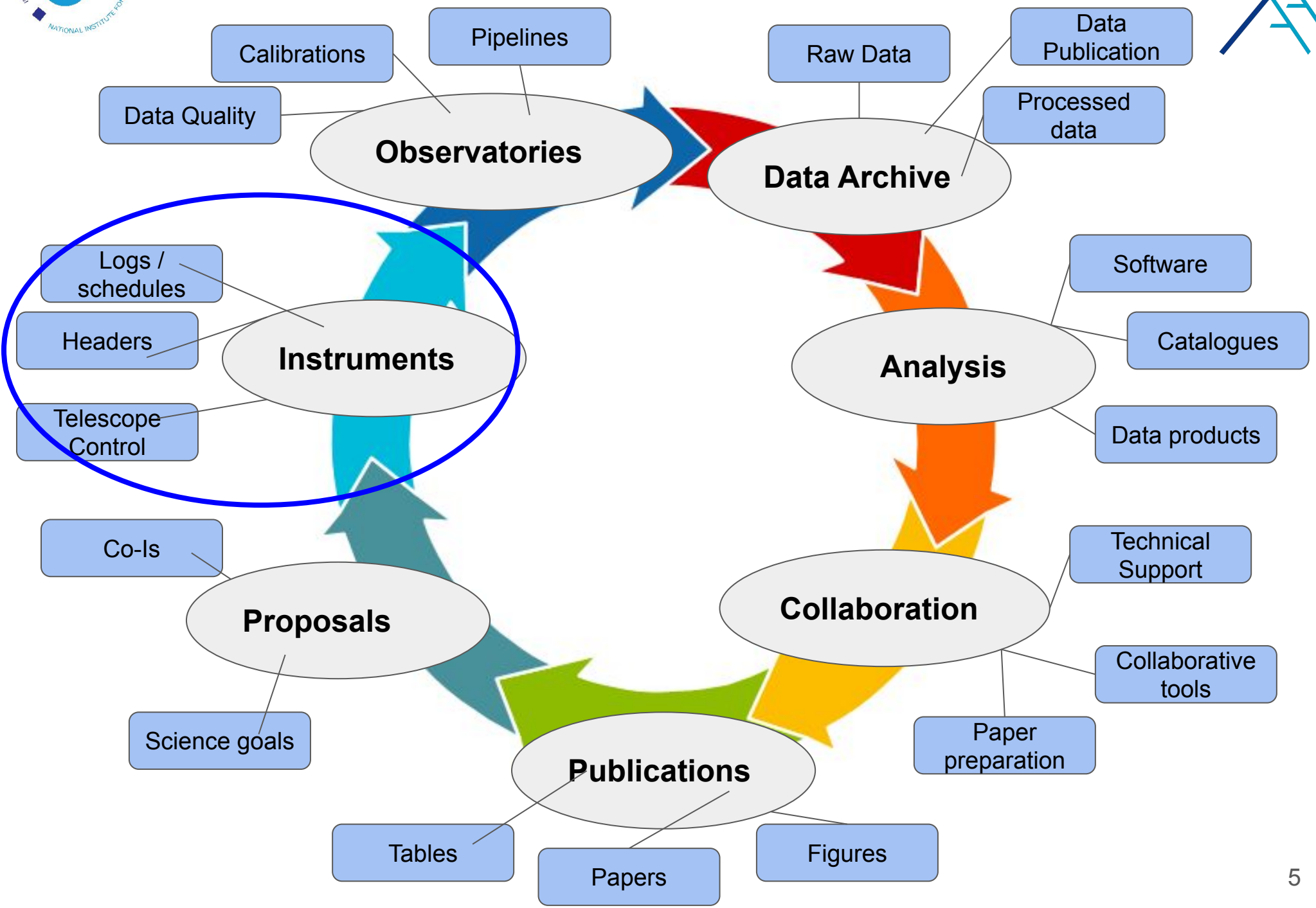




Do not forget to **REGISTER** yourself -whenever possible- in advance, it will allow you to **easily access** your **observation preparation and data**!

https://sso.ia2.inaf.it/home/

Astronomical Data Life Cycle



Raw Data: the scientific and technical content

Observations are composed by several components:

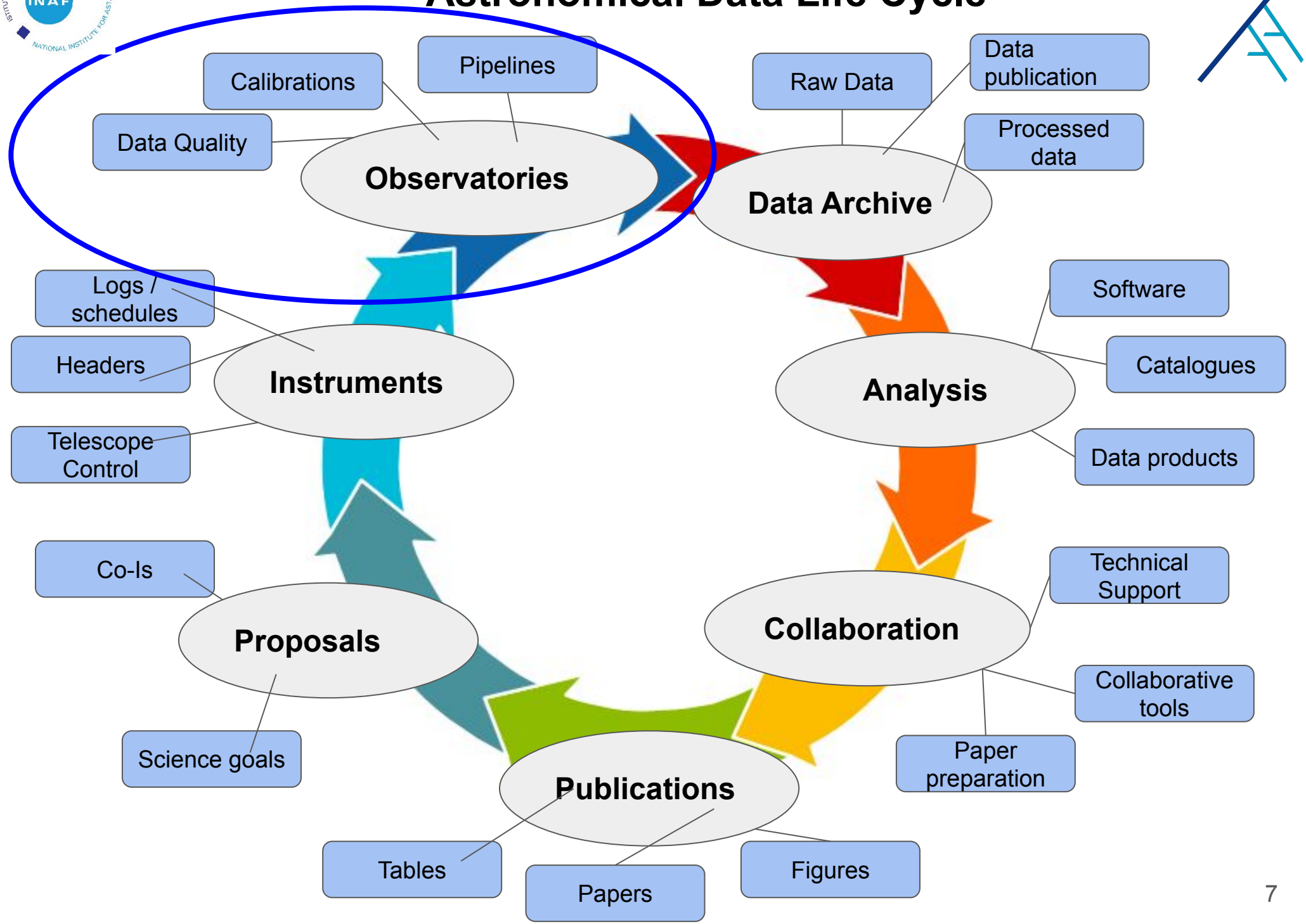
- **the schedule**
- **the scientific exposures**
- **the image/spectra description (metadata)**
- **the calibration exposures**
- **the engineering data**
- **the night shifts**

All those components mixed together contribute to the correct data analysis and interpretation.
Having all the components available and interoperable thru different systems allow **data reuse**, preservation and curation.

- **Continuous interaction with Data Providers to evolve the data models**
  - Telescope staff, instrumentation specialists, support astronomers, technical staff etc..
- **continuous interaction with standards experts**
  - IVOA, RDA, FAIR,

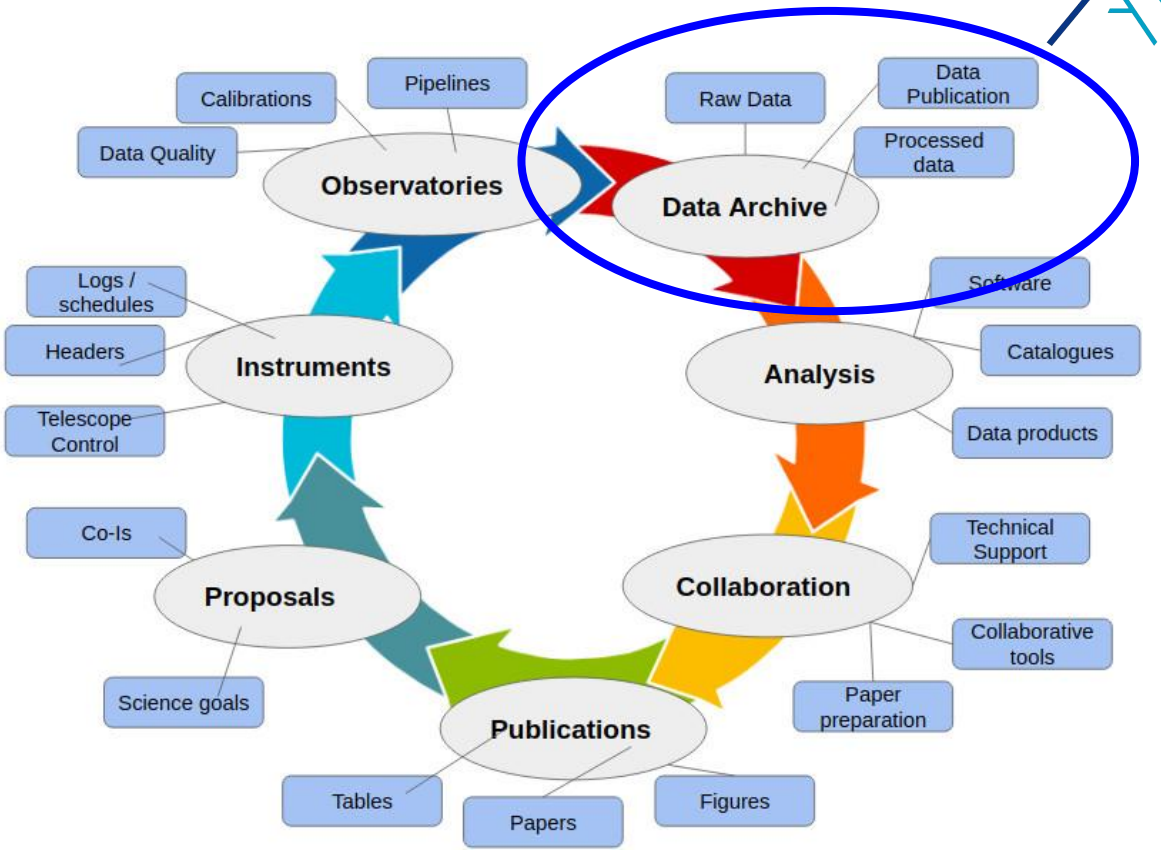
See and remember the **FAIR** principles!

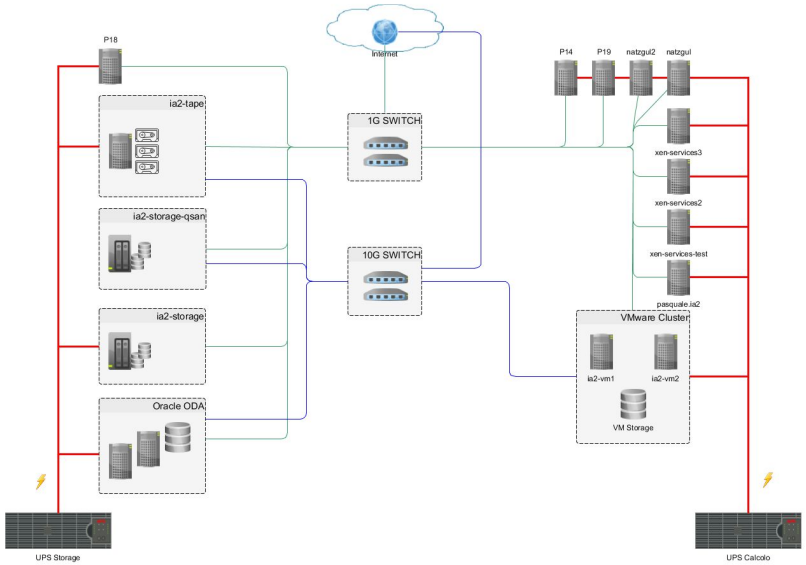
Astronomical Data Life Cycle



From Observation to data products

- store data for long term preservation;
- store data in a repository for temporary store;
- store data in <u>on line</u> archive;
- serve data in VO compliant manner;
- allow operations on data in a user space;
- allow computation on data;
- allow for interactive/collaborative tools operations on data;
- DOI……..





- **Hardware IA2 @ TS**:
  - 800 TB
    - 300 used + 400 free TB on line
    - backup : 100 TB for VMs
  - T950 HPE LTO-8 of 1.25 PB expandible to 12.5 PB (coming next month)
- **Hardware @ other sites**:
  - IRA : 60 TB on new machine (x testing)
  - SRT : 60 TB on new machine (buffer to Ts)
- **Hardware owned by others**:
  - IRA : 40 TB Radio Distributed Archive
  - SRT : 1 TB (pulsar testing machine)
  - Serra La Nave : 500 GB on site
  - LBT : 12 TB upgraded 1TB /y Full LBT Archive
  - Asiago : 500 GB on site
- **Bandwidth**: 10Gb/s GARR

From data providers to science

Data Providers: Telescopes, Projects (Simulations or Surveys), Users

Types of data:

- Structured data: well organized data following defined standards and always containing all the descriptors to guarantee the FAIR principles;
  - Easy to handle by a automatic routine, can be ingested into a searchable archive;
- Partially structured data: data customized to fulfill specific needs but not always coherent with the previously produced one. Sometimes it follows a self define standard and can be seen as an evolving data model dataset;
  - Require frequent updates to the data model and can be difficulty be published in a unique catalogue;
- Shared data: fully incoherent data composed usually of a mix of data models and formats. Totally customised to the current utilization and defined by human approach (no standardization). It usually contains a mixture of raw, reduced data and lists of parameters, comments, wiki pages etc;
  - require a collaborative tool to support data sharing and cannot be published by an archiving service. It can be part of the DOI publication but in any case presents difficulties in data searching.

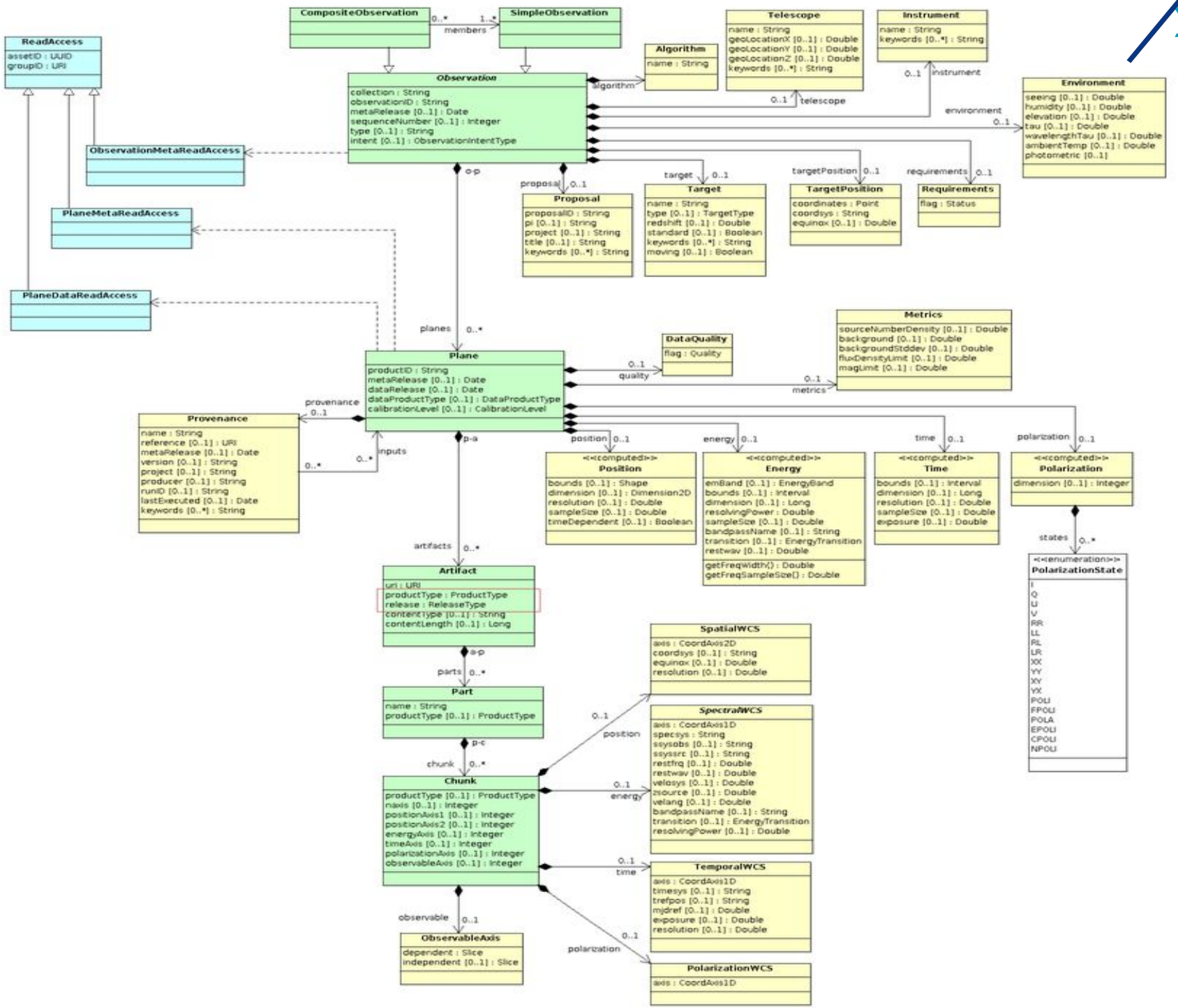
Structured Data: what is a data collection?

- Data collection is the **systematic approach** to gathering and measuring information from a variety of sources to get a **complete and accurate** picture of an area of interest in a **standardized and established manner** that enables the collector to **answer or test** hypothesis and evaluate outcomes of the particular collection. The goal for all data collection is to capture quality evidence that allows analysis to lead to the formulation of convincing and credible answers to the questions that have been posed.
- Accurate data collection is essential to maintaining the integrity of research, guaranteeing findability, accessibility, interoperability and reproducibility (FAIR principles);
- Impact of faulty data:
  - Inability to answer research questions accurately;
  - Inability to repeat and validate the study.

What is a data collection?

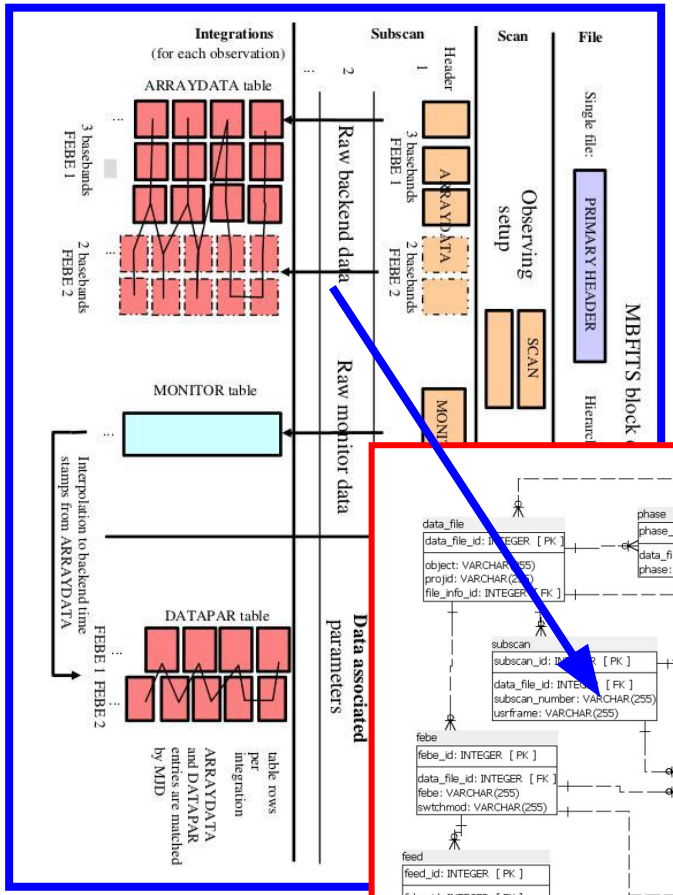
- A collection (used as a noun) is the topmost container for grouping related documents, data models, and datasets.
  - **Data model**: is an abstract model that organizes elements of data and standardizes how they relate to one another and to properties of the real world entities;
  - **Dataset**: is a collection of data. Most commonly a data set corresponds to the contents of a single database table, or a single statistical data matrix, where every column of the table represents a particular variable, and each row corresponds to a given member of the data set in question. The data set lists values for each of the variables, such as height and weight of an object, for each member of the data set. Each value is known as a datum. The data set may comprise data for one or more members, corresponding to the number of rows.

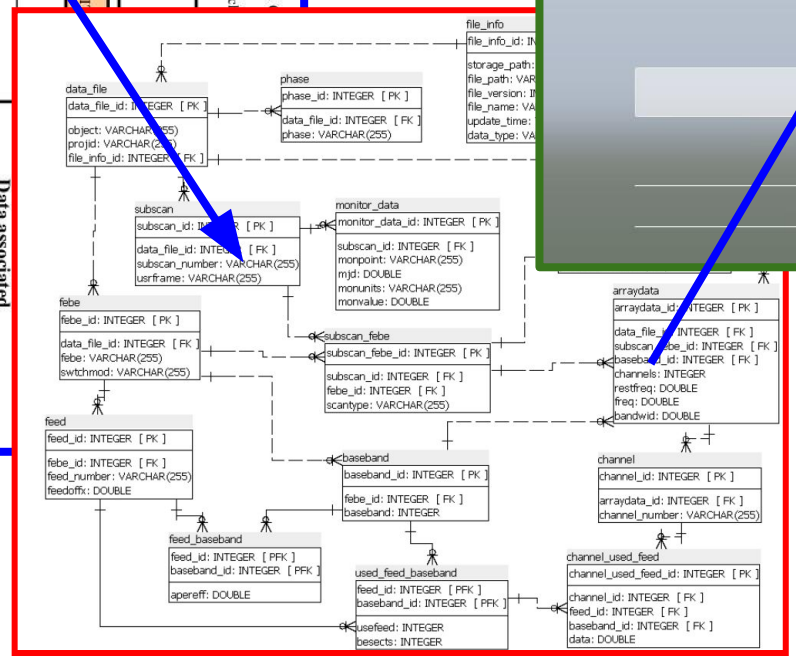
Data models: ObsCore and CAOM



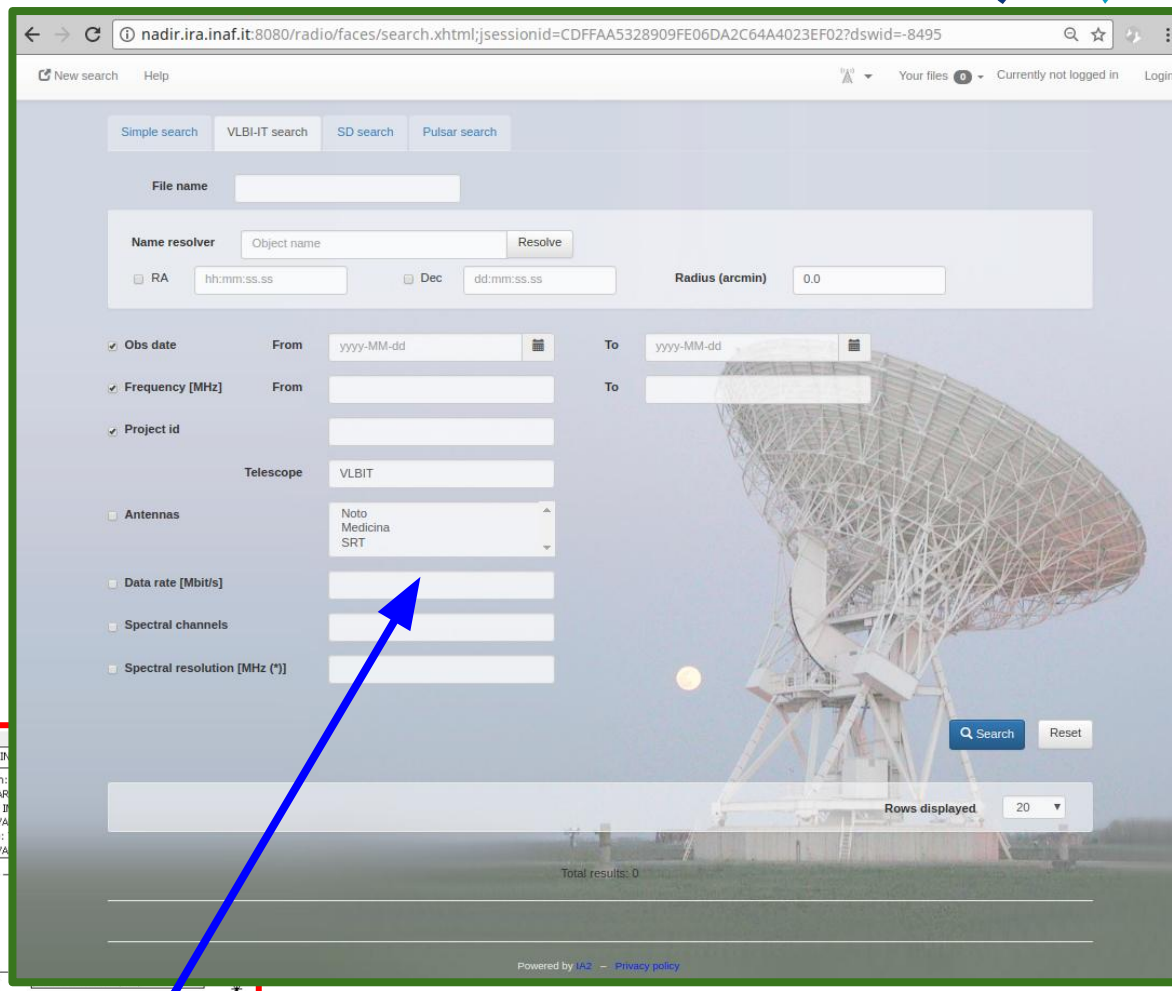
Data Models to Archives

Standard





Data Model

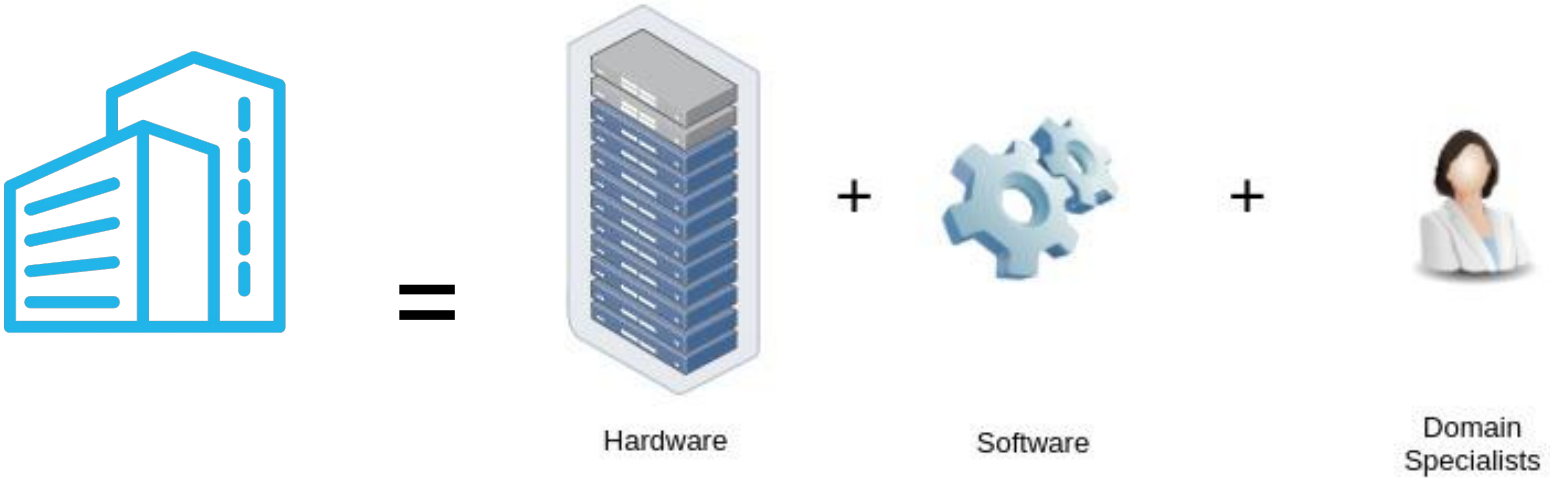


Querable dataset

IA2 support to Science

- **Store and preserve** astronomical data (observed or simulated);
- **Support data providers** in correctly set up Archives:
  - from raw data to calibrated one from Telescopes;
  - simulated data (exoclimates, intrigoss, cosmological)
- **Publish public data** through the VO services;
- **Support Astronomers in data retrieval** via web interfaces:
  - search on public data without login;
  - after login, a list of proprietary datasets are presented and filters can be applied;
  - using filters to find data;
  - single file direct download;
  - user space for bunch of files;
  - possibility to download VOTables of queried data;
  - possibility to download a CSV file of queried data;
  - possibility to download a URL list of files to download them using external tools like WGET;
  - name resolver to find coordinates of objects;
  - connection via SAMP HUB to link VO clients like Aladin or Topcat;

Necessities: user support

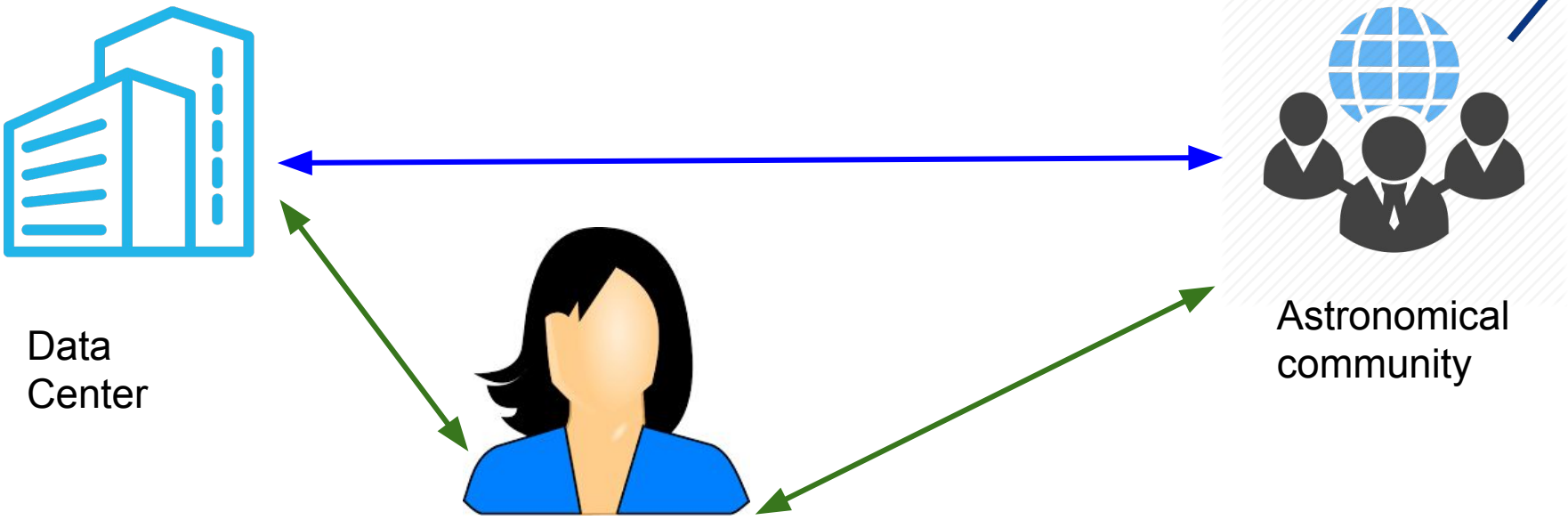


Data Center

The acquired know how from data provider side made evident the collaboration between domain experts and scientific community is essential to correctly identify data models and to define the specific content of the exposed services. In this scenario, Virtual Observatory, FAIR policies, Open Data, RDA play a fundamental role.

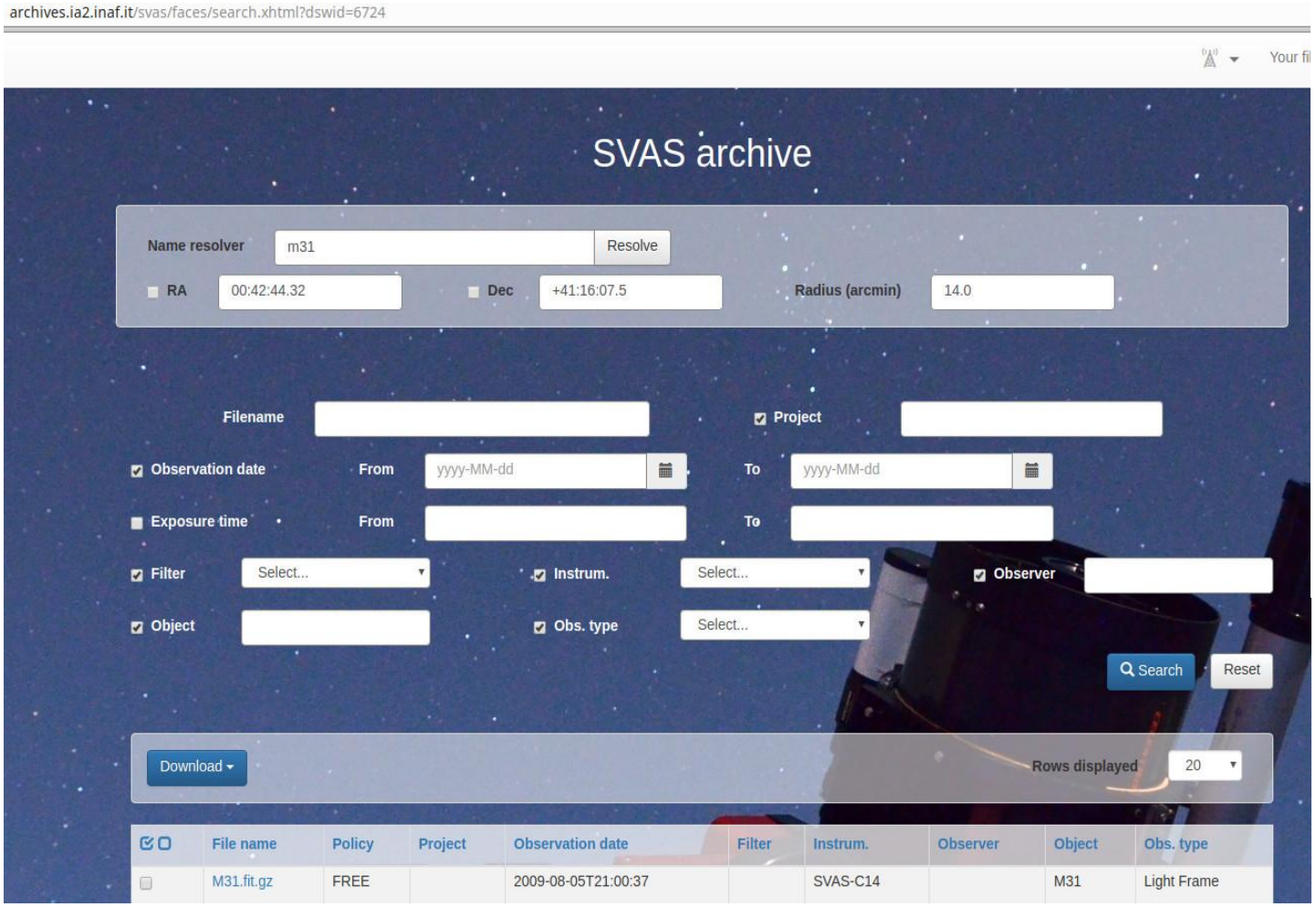
User support for data retrieval is essential too!!

Necessities: user support



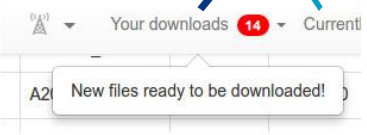
- ARC and AENEAS experience: for the future 2020 era telescopes a scientific support to extract the correct information from the huge amount of data is ESSENTIAL;
- CADC experience: user space available to observing program collaborators where to manage and handle data with available or provided pipelines is highly desirable in order also to interoperate and cross correlate different datasets and scientific products;
- DOI: papers and datasets!! After paper publication, datasets should be public!

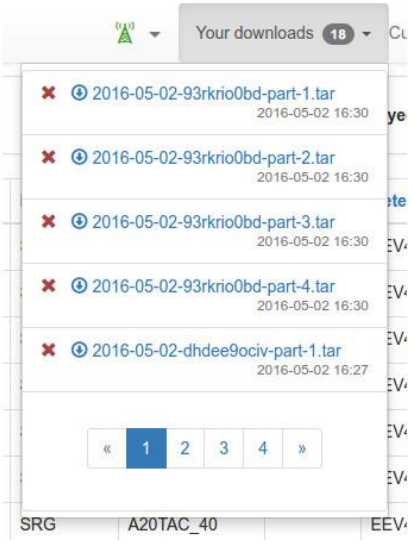
IA2 web interfaces



http://archives.ia2.inaf.it/tng/faces/help.xhtml

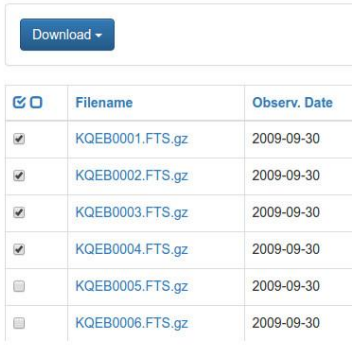
"Your files" menu





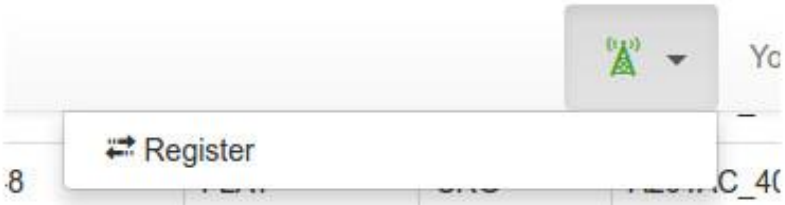
Download multiple files

As tar archives

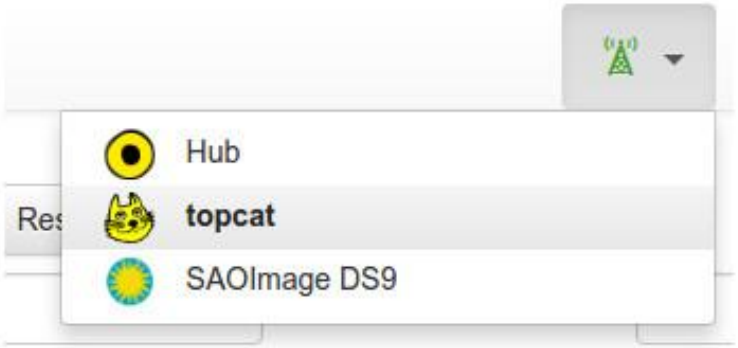


Using external tools (e.g. wget)

IA2 SAMP



If your SAMP hub is running the icon is green and you can register on it to see the other applications.



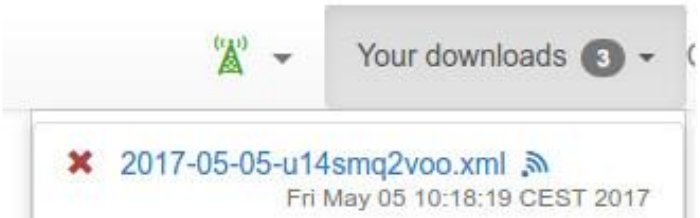
When you are registered to the hub you can select the target application.

If you are unregistered or you haven't select any application your messages will be sent to all active applications.

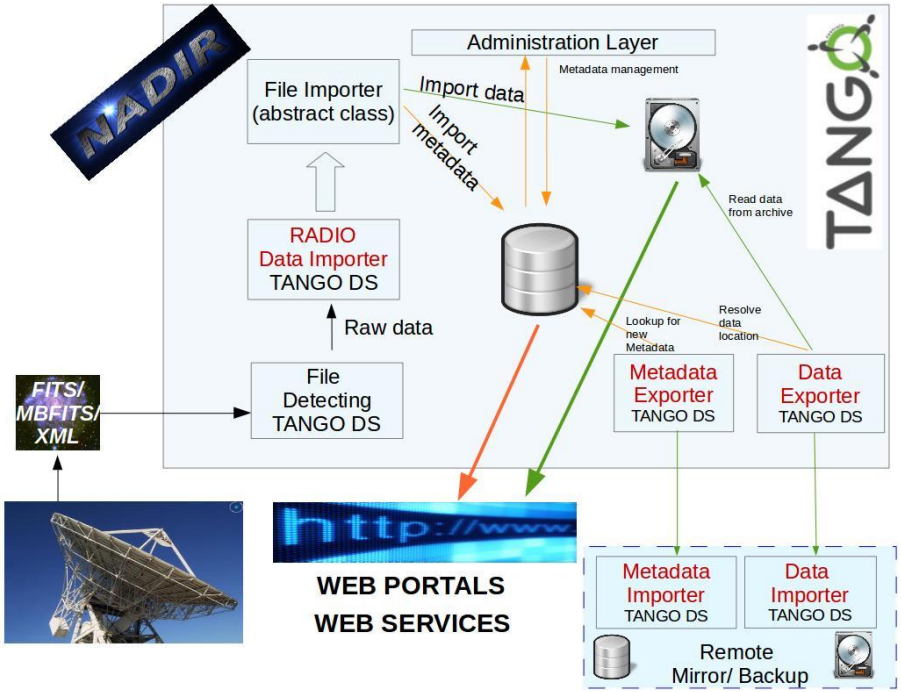
You can:

- Send a VOTable by SAMP (MType `table.load.votable`)
- Send an image by SAMP (MType `image.load.fits`)

When a SAMP hub is detected, a "signal" icon will appear near search result files and generated VOTable files.

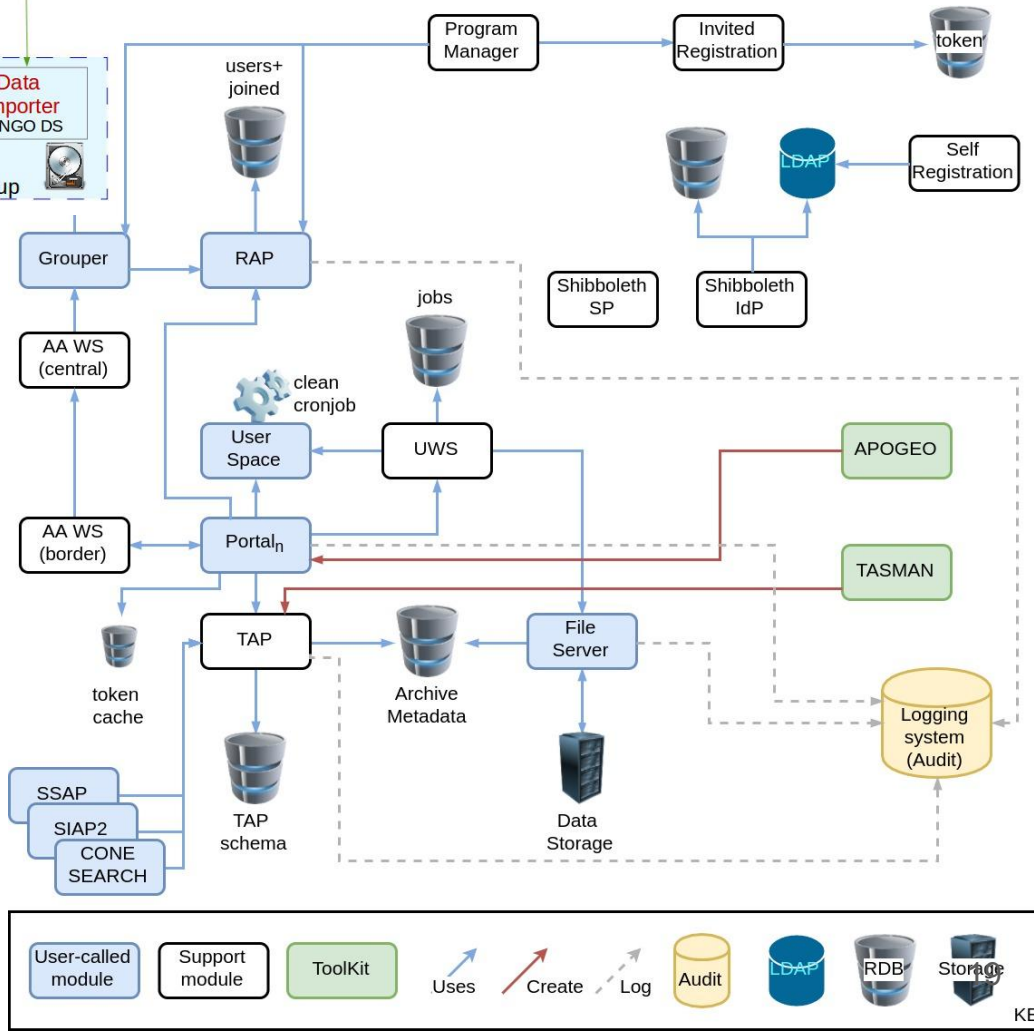


IA2 services

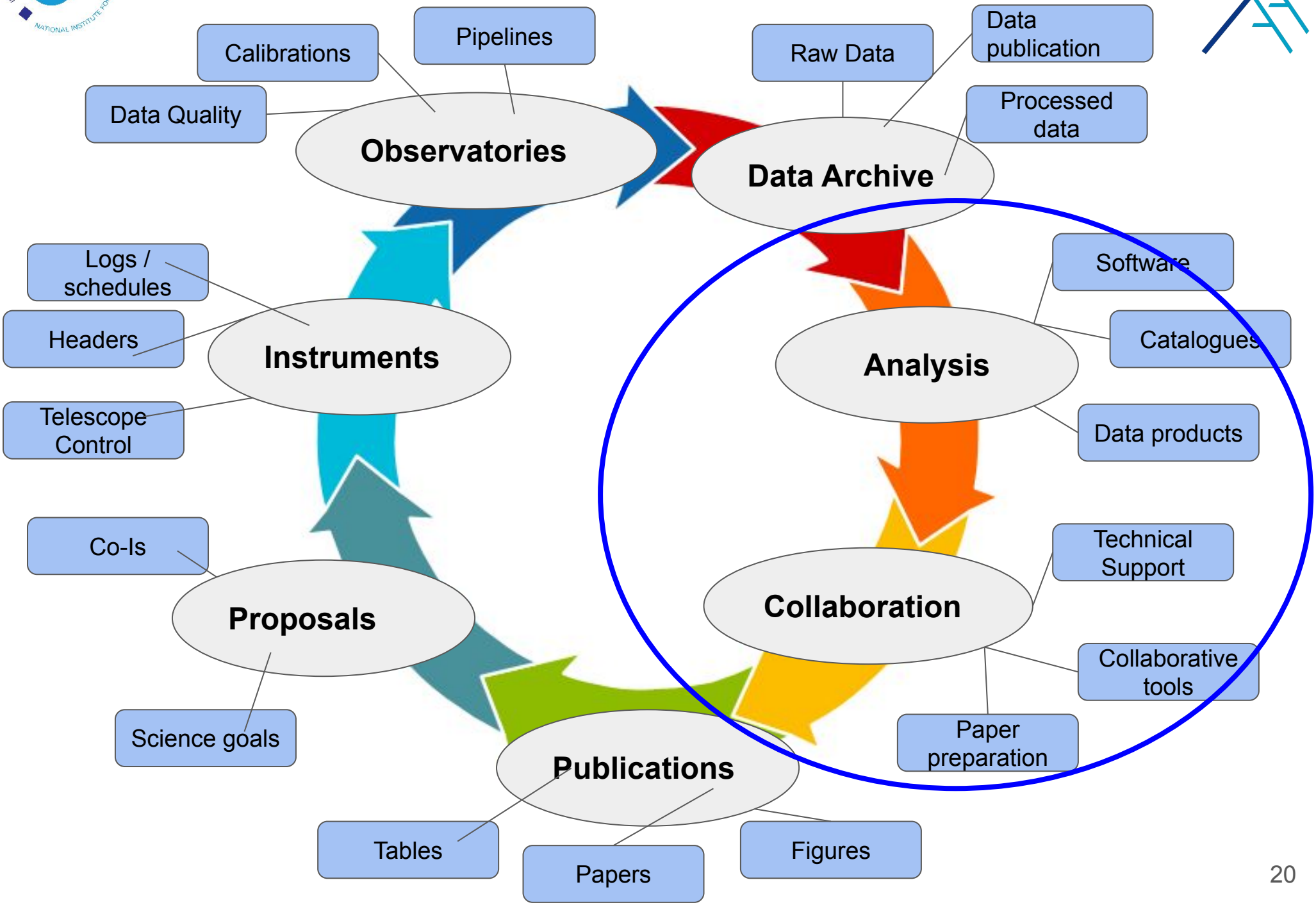


NADIR: Ingestion software

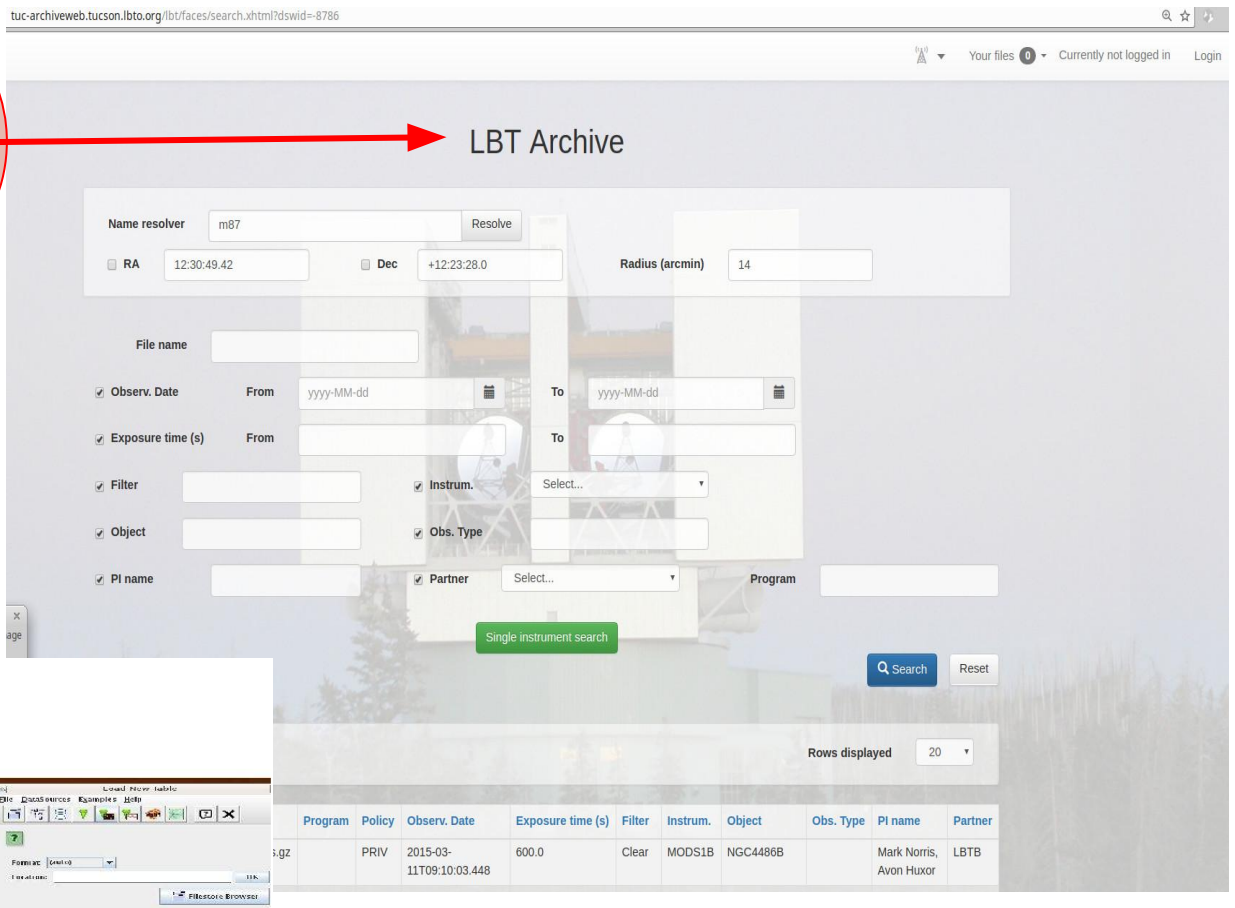
Web interfaces and services and authentication and authorization software



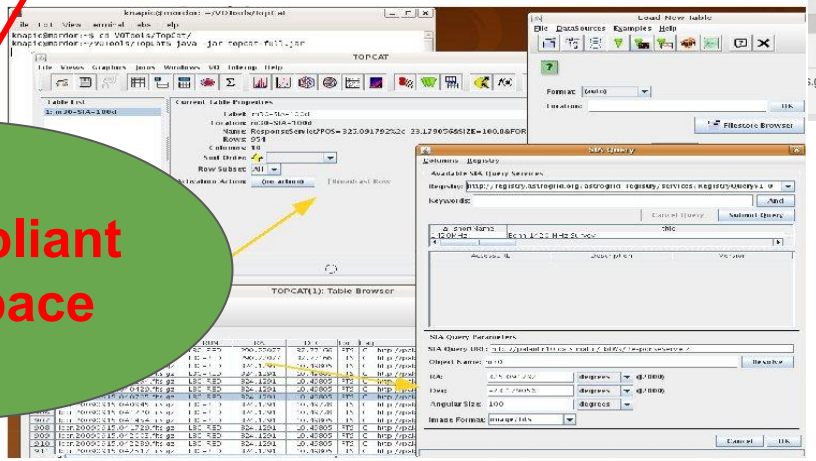
Astronomical Data Life Cycle



Login to access tools

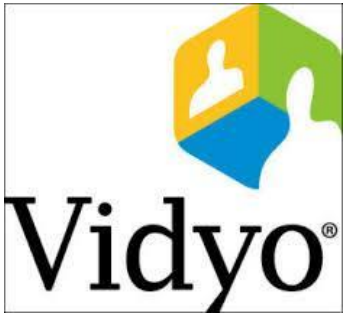


VO Service & TopCat



VO compliant User Space

WEB: Services, portals e VO compliancy



WEB: Services, portals and VO compliancy

Science ready data: the TNG example

- ➔ reduced data using pipeline DRS published on TNG portal following observing program based policy (from 1 to 3 years);
- ➔ all HarpsN and GIANOB observers have the possibility to reprocess data using in an interactive way the installed pipelines (DRS and Gofio). Pipelines are updated on YABI platform by IA2 team.

Yabi : platform for the management of data reduction pipelines.

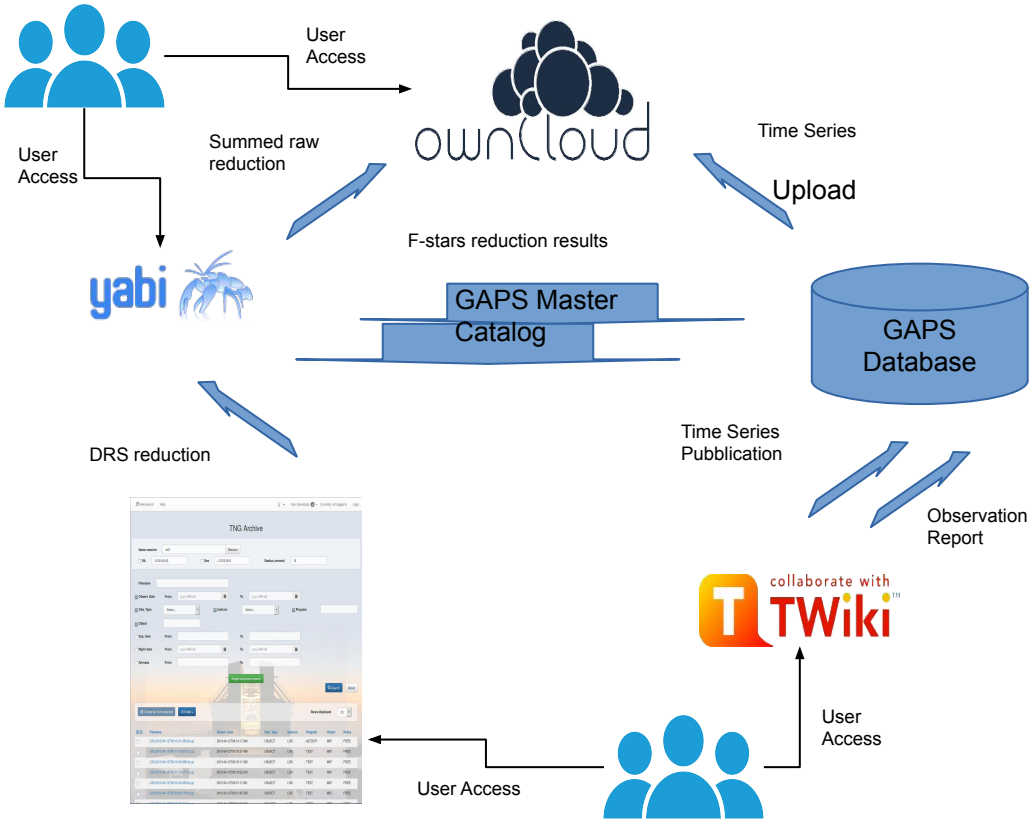
Advantages:

- ➔ software up-to-dated to the last pipeline release;
- ➔ private pipelines, potentially public data;
- ➔ intuitive and interactive pipeline usage;
- ➔ users do not need computing power.

Disvantages:

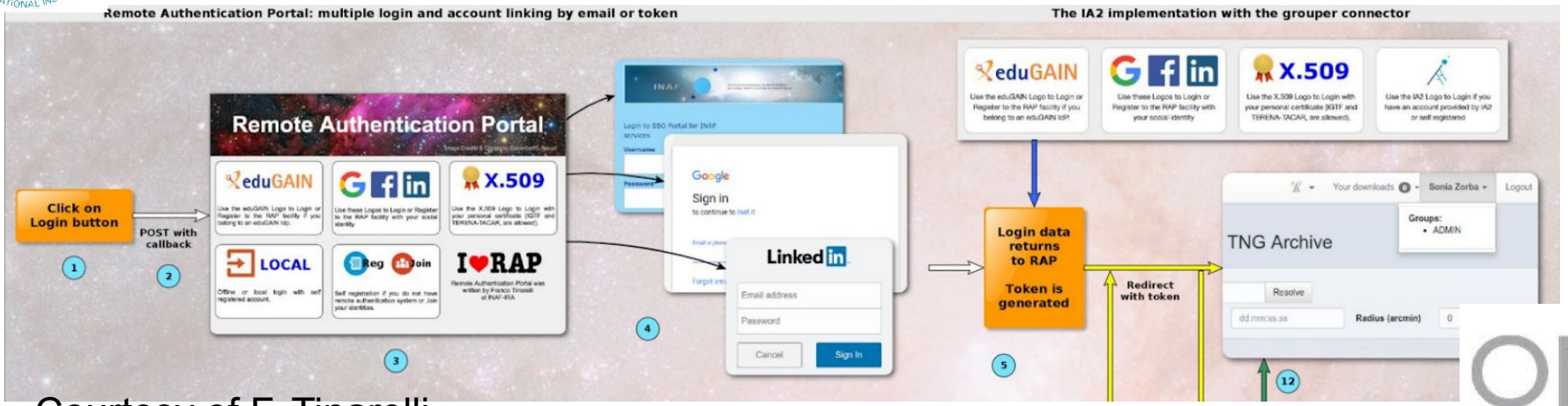
- ➔ local data management (currently no reingestion of reduced data for privacy reasons)
- ➔ currently not integrated with RAP

Computing platforms



- IA2 manages data in a distributed manner on 3 continents!!
- GAPS experience bring important know-how. Pipelines and workflow management systems will be the must of 2020 Era Telescopes;
- IA2 allow state-of-the-art authentication and authorization mechanisms. Same results will be applied in pilot project for SKA and AENEAS.

Remote Authentication Portal

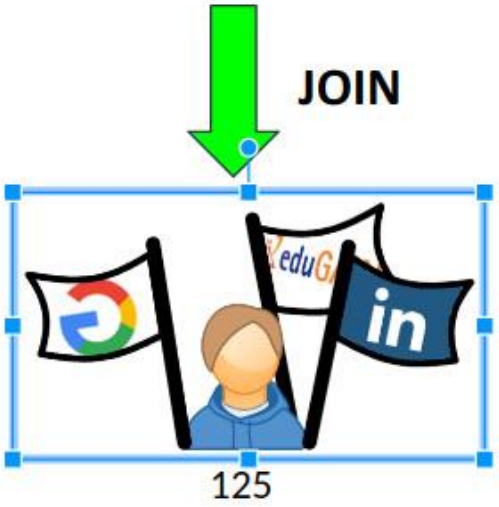


Courtesy of F. Tinarelli

+

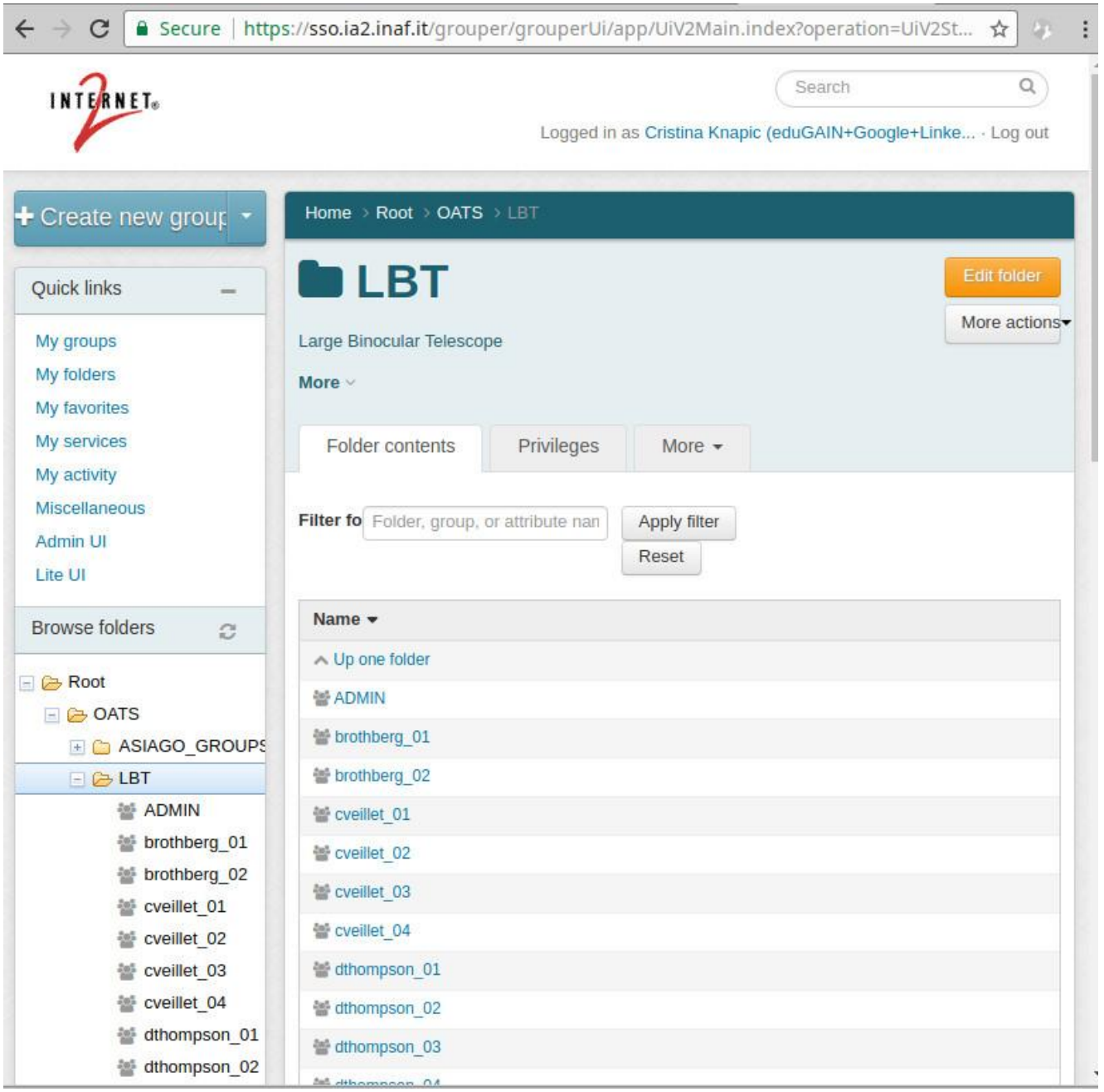
ORCID
Connecting Research and Researchers

The same person, multiple identities, multiple user ID



One user, multiple identities, the same user ID

Grouper

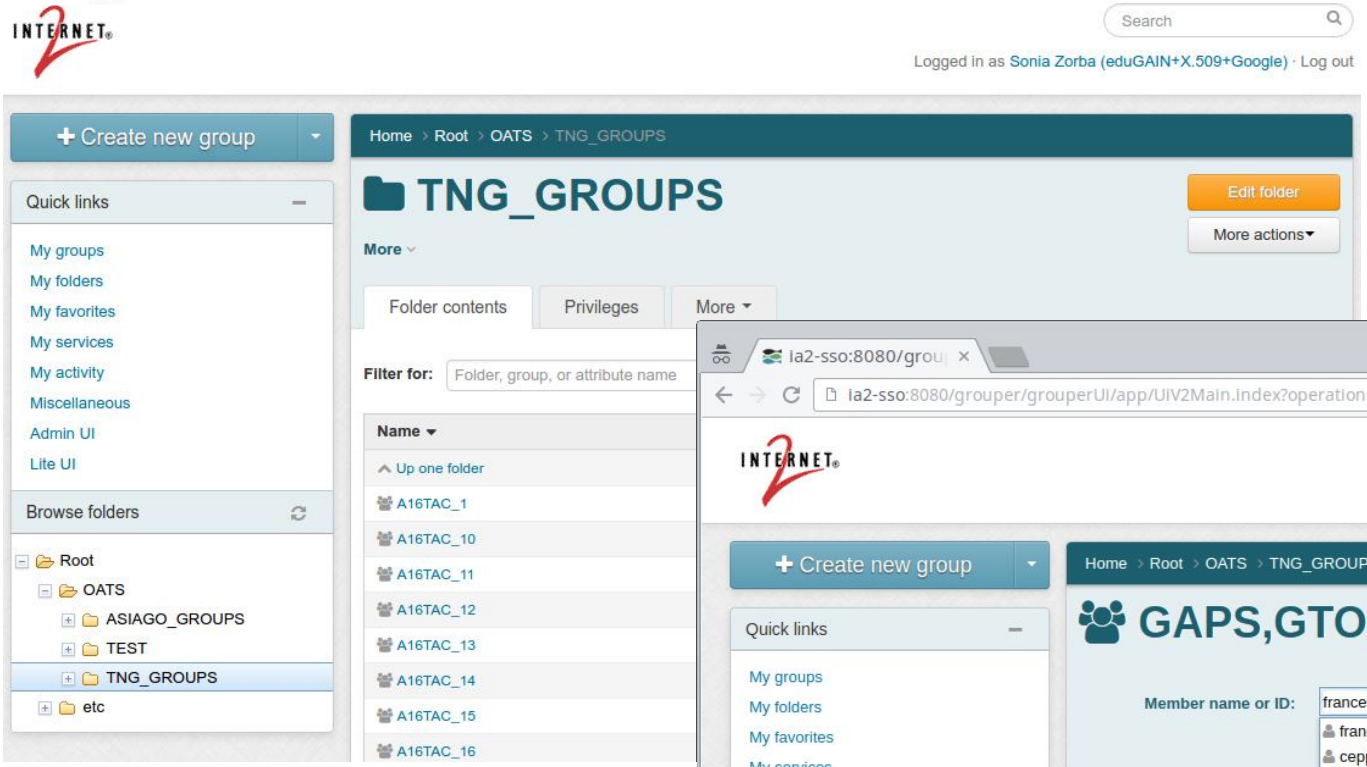


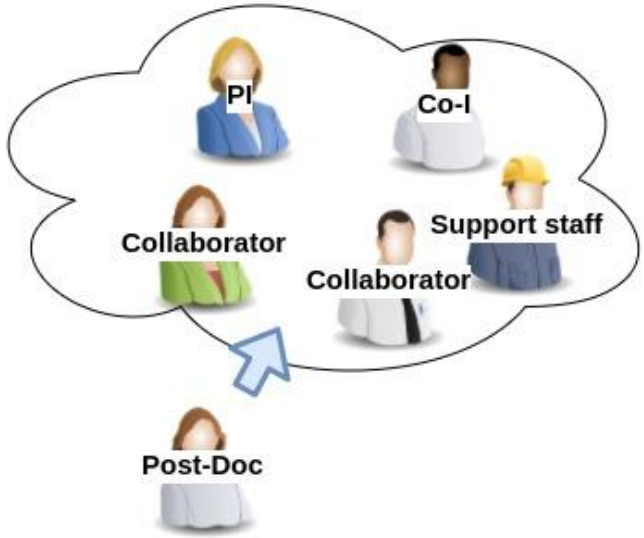
You can give access to a set of data adding a member to your group (corresponding to a program in this case).
More than one application could query on the group management system and allow different operation and privileges to the members, so you can easily collaborate.

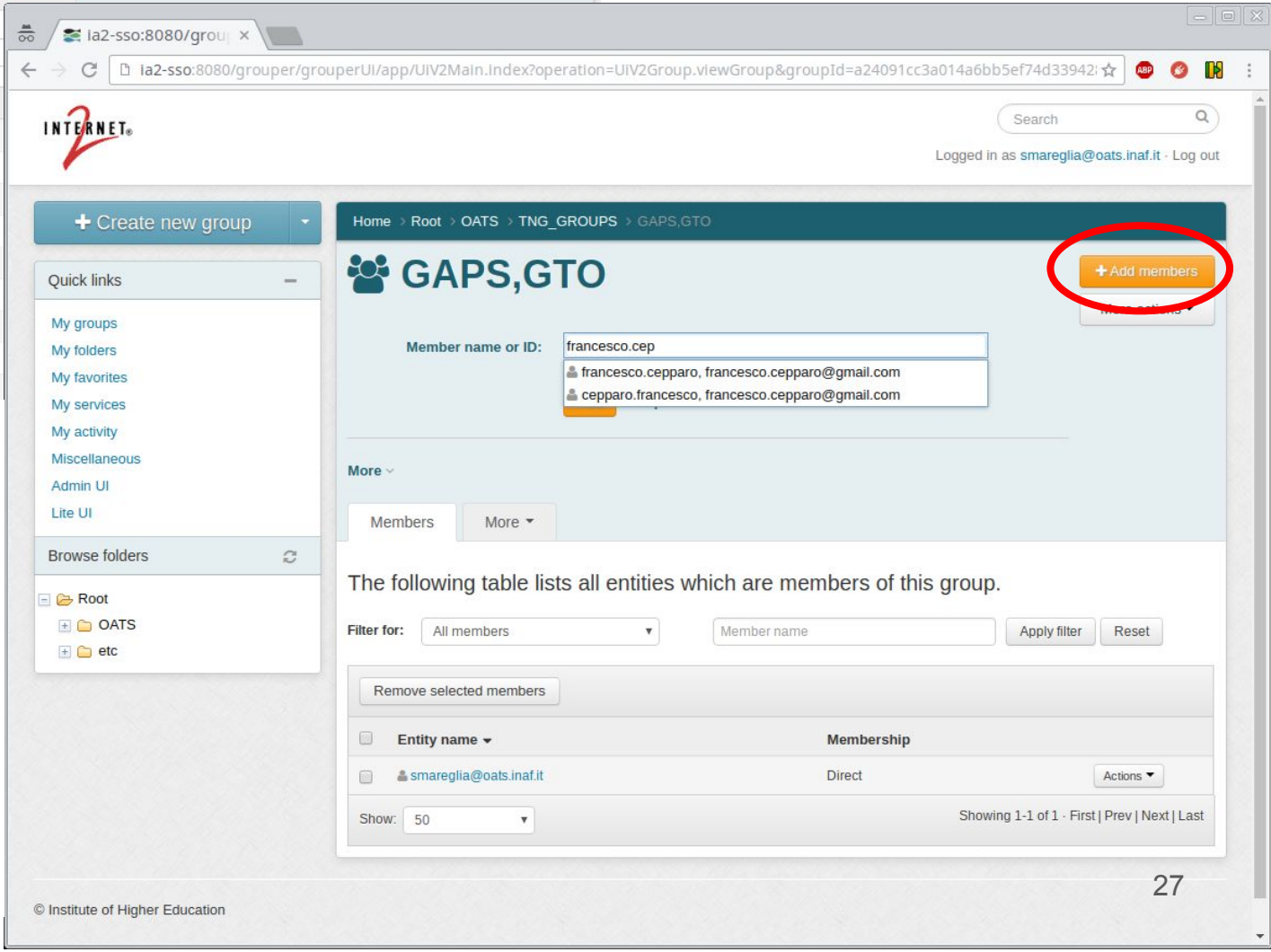
To run a specific temporary project, a user space is a very nice place where to share files and run applications that cannot be operated locally, but needs more resources than our laptop could offer. It is in development…..

Adding a user to your group

https://sso.ia2.inaf.it/help-grouper/







Under development: VOSpace for storage and computation

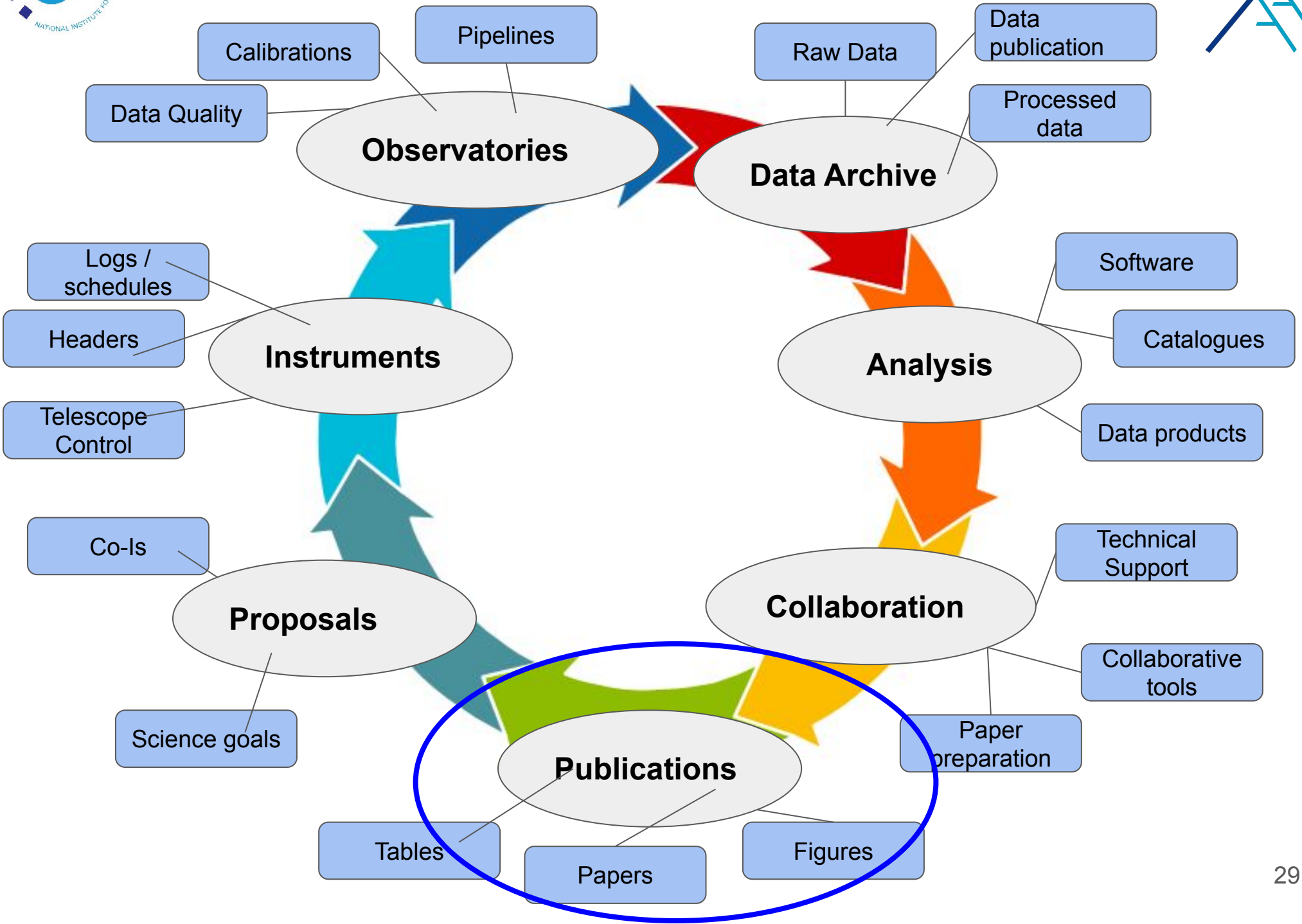
VOSpace implementation compatible with CADC implementation, possible share of authorizations (thanks to S. Bertocco, G.Taffoni, S. Gaudet, P. Douler, B. Major experiments within EgiEngage)

Two levels of computation:

1. user approach to interactive pipeline with no HPC/HTC and small data volumes
   a. RAP + Yabi;
   b. Containers on IA2 infrastructure;
2. user approach using bash processing with HPC/HTC needs
   a. Containers or VMs on Chipp;
   b. GCloud.

Processing close to Data
+ POCs

Astronomical Data Life Cycle



Your data

- Remember your reduced data set is composed by several components:
  - raw and reduced datasets (see above)
  - Pipeline
  - reduction parameters (filters applied..etc..) and intermediate products
  - tables / plots

DOI for Data Sets → supported by ICT
What about creating catalogues of DOIs using additional metadata (data descriptors)? It allows datasets exploitation and burst the number of citations

See details in R. Smareglia talk

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.

The ideal plan

- Improve the services offered (link to proposal submission);
- Increase the number of services offered (possibility to automatic **re-processing of** data with last version pipeline, see ARI-L project; basic **Pipelines** for all the instruments);
- provide data curation before publication;
- Increase the number of supported data models if not compatible with CAOM in all the chain (archives, automatic data reduction software, VO);
- Grant access to all the offered services via SSO;
- Adopt a VOSpace compliant service to allow **user space** utilization (storage + computation);
- data/catalogues publication increasing DOI metadata;
- **Improve the user experience in order to stimulate the Astronomical Data life cycle**.

Conclusions

- Target of Data Centers:
  - Support science offering Archival services, computing power and maintenance of both

- Target of the Astronomers:
  - Use of data center services
  - Have their own User Space where to make science!
  - share data, papers, ideas, everything!!
  - Reporting problems
  - Suggest upgrade/necessities

Thanks for your attention!