LBT – Distributed Archive: Status and Prospective

R. Smareglia , C. Kanpic, M. De Marco
and the IA2 Team

LBTO 2014 Users' meeting – 23 March 2014

LBT-DA: Status

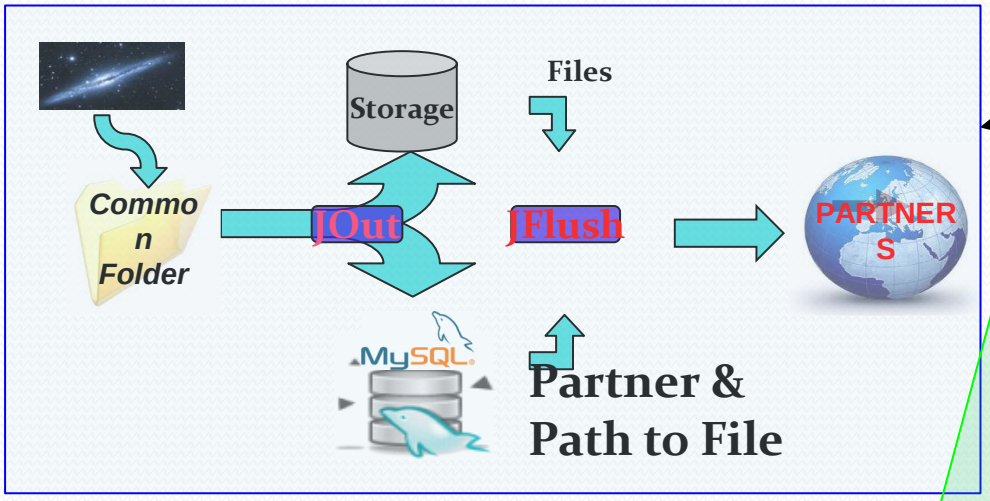
- LBT-DA is FULL managed from IA2 (INAF) Data archive center since 2005

➔ http://lbtarchive.as.arizona.edu

- Centro Italiano Archivi Astronomici (IA2) main goals are :
  - Archiving systems
  - data curation and preservation,
  - distribution over several geographical distributed sites,
  - providing services and tools (TWiki, work-flow, etc..)
  - data publication in the VO

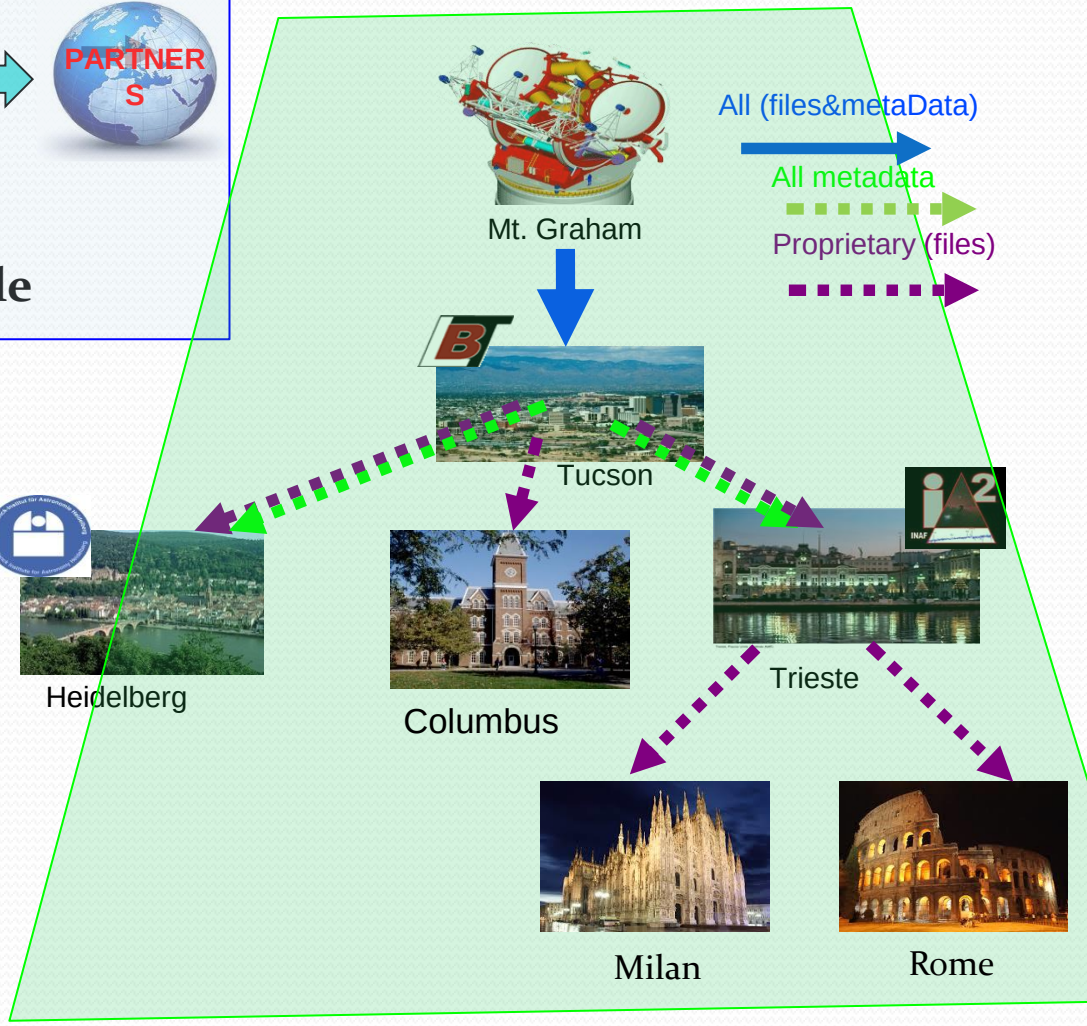
of *Astronomical Data*

LBT-DA @ IA2



OAIS standard implementation

Data distribution policy



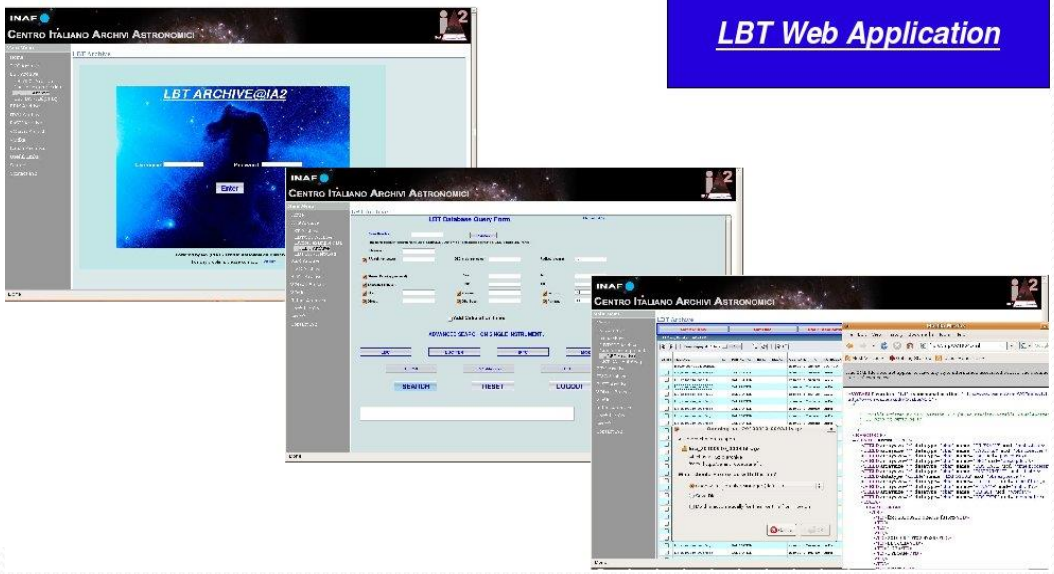
LBT-DA: few numbers

- Speed data transfer:
  - *Data are ingested on the mountain after few seconds from when present on disk*
  - *Move to Tucson in less than 20-30 sec.*
  - *Move to Heidelberg or Trieste less than 3-4 mins.*
- *~ 10 TB of compressed raw data (Tucson archive)*
- *465.000 files (118.000 calibration)*

LBT-DA: few numbers 2

- Metadata and calibration data are public;
- Data are accessible only by PI or PARTNER
  - (only INAF have a data policy : 1 year for raw data)
- *Access since 2010 (status at 2014 03-19):*

Web interface statistics:					
		LBTO	LBTB	INAF	
				Web	Vo
last week:	n of access:	66	0	197	606
	n of downlaod:	6	0	156	148
last month:	n of access:	1264	36	1272	2414
	n of downlaod:	268	13	1189	1175
Total:	n of access:	21250	4123	30611	78876
	n of downlaod:	5052	1328	25750	24940



IA2 Archives new Project:

NADIR

Issues to solve:

- Meta-data information could change in:
  - Contents;
  - Format types;
  - Keywords.
- Data formats:
  - Extensive use of ASCII and PH for calibrated data
  - FITS is not the only astronomical format (HDF5, PDS, MBFITS etc..)
  - Standards are evolving
- Consistency in meta-data content both into the file and database in case of value correction
- Data models should be, as much as possible, code independent
- Code re-usability and configurability
- Scalability in both serial and parallel ways
- Consistency over several distributed archives and secure differentiation.

NADIR is one configurable and flexible software that *answer the challenging problem of archiving software reuse and scalability*. It can handle also calibrated data.

NADIR's offered solutions

- Modular software, optimized to be as much as possible flexible;
- Scalability in both serial and parallel data distribution paradigms;
- **Handling of calibrated data**
- **Policy and versions revised** easily, in a flexible manner;
- Consistency on geographically distributed archives and secure differentiation in archives content.
- Strong logging and error handling;
- Possibility to correct meta-data content, maintaining consistences across distributed archives;

NADIR Mandatory Requirements:

- **INSTRUMENT;**
- **OBS DATE;**

NADIR functional requirements:

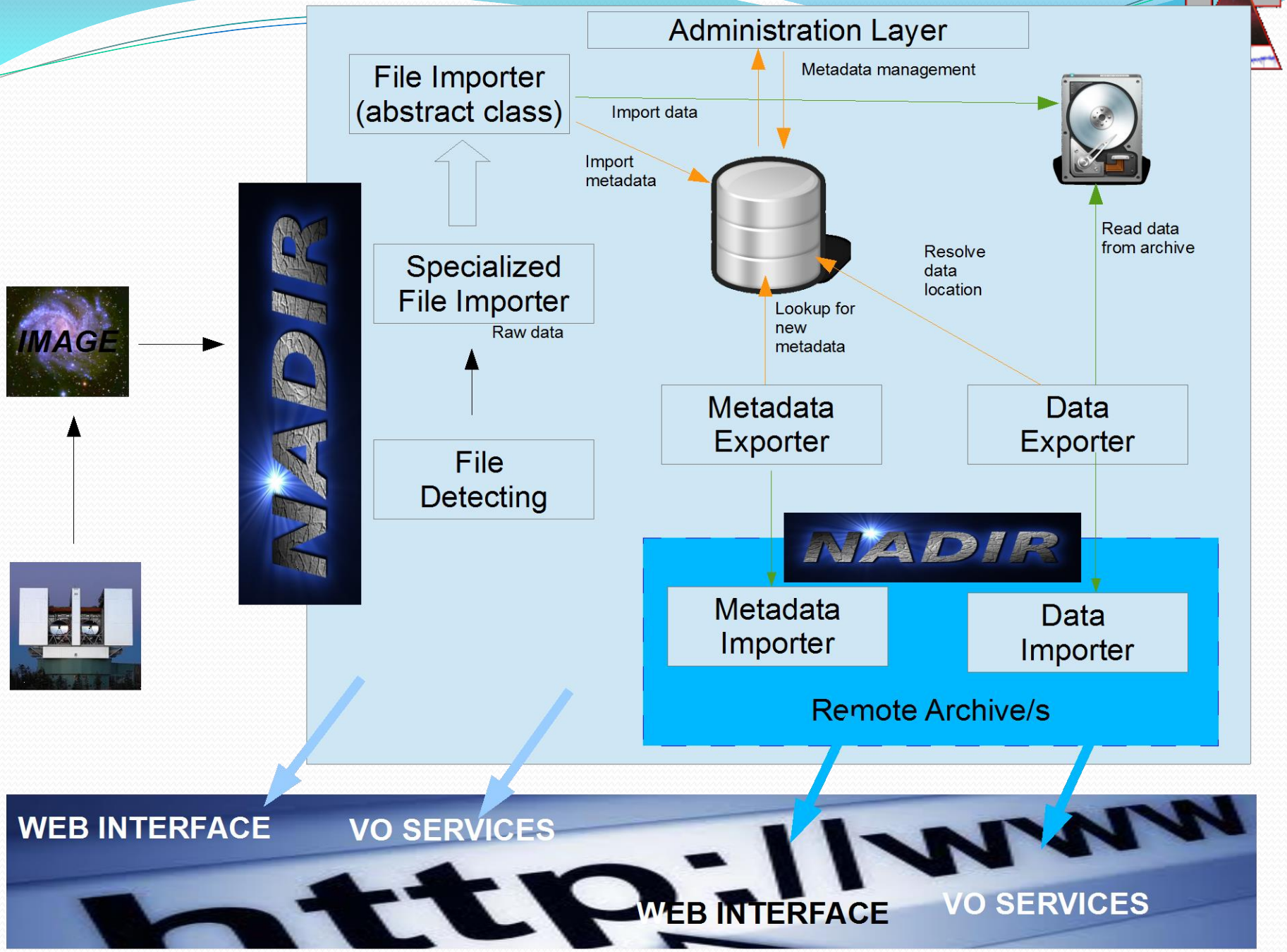
- **PARTNER;**
- **PINAME;**

NADIR non functional requirements:

- **Coherent filling of fits keyword values in terms of types and values consistencies to allow query efficiency;**

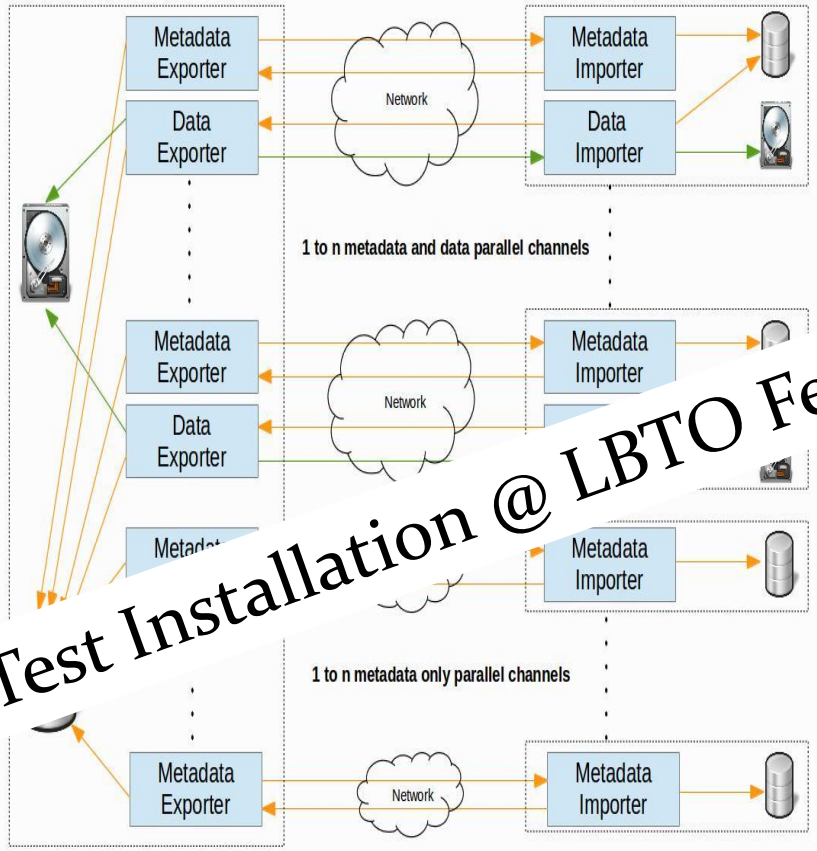
Data propagation depends on ingestion date and policy. Policy depends on OWNERSHIP.

No ownership = no data distribution!

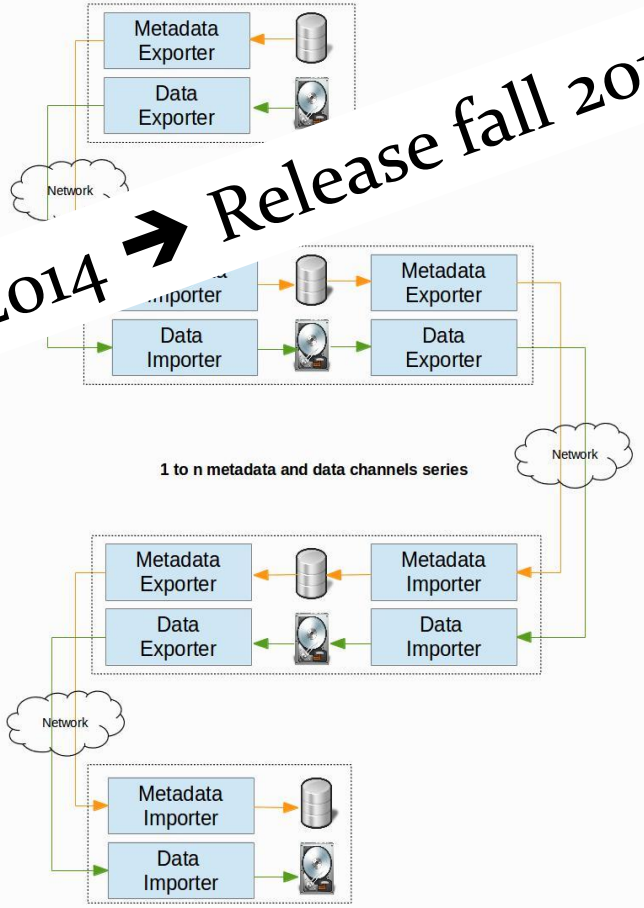

Administration Layer
File Importer (abstract class)
Metadata management
Import data
Import metadata
Read data from archive
Resolve data location
Lookup for new metadata
Specialized File Importer
Raw data
File Detecting
IMAGE
NADIR
Metadata Exporter
Data Exporter
NADIR
Metadata Importer
Data Importer
Remote Archive/s
WEB INTERFACE
VO SERVICES
WEB INTERFACE
VO SERVICES

Meta-data and data distribution

Parallel distribution



Serial distribution



Test Installation @ LBTO Febbrary 2014 → Release fall 2014

But an archive is not only that...
You want more from your data!

Archive means having raw, science ready products (calibrated data, catalogs and so on..) possibly VO compliant and/or accessible by VO services (it doesn't mater if the archives are local or remote). So:

* **data interfaces**
* **VO services**
* **work-flows / pipelines**
* **data curation and preservation (DOI)**

The archive front end is a Science Gateway!

But an archive is not only that...
You want more from your data!

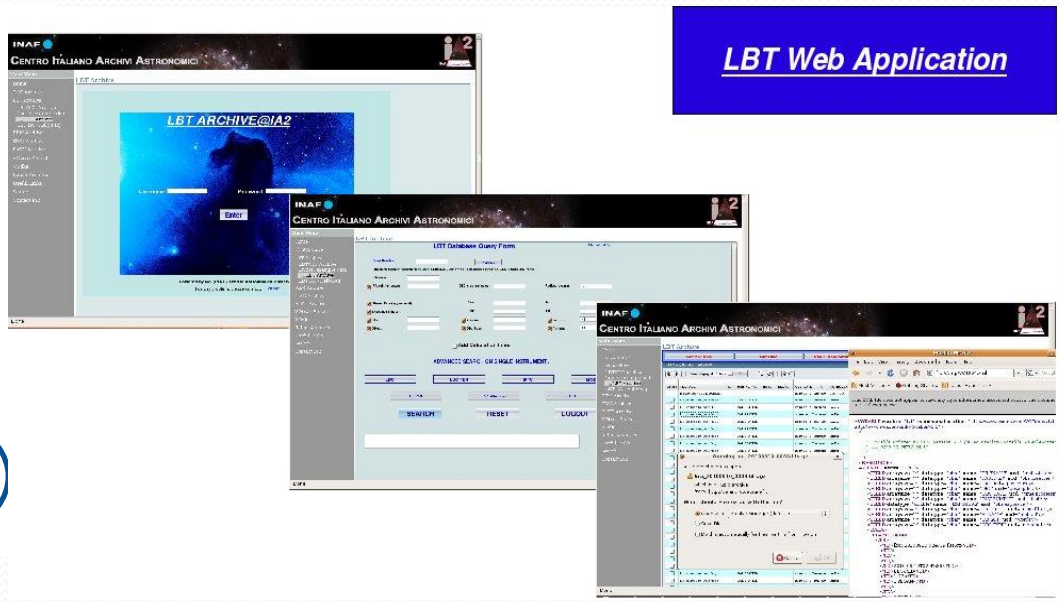
Archive means having raw, science ready products (calibrated data, catalogs and so on..) possibly VO compliant and/or accessible by VO services (it doesn't mater if the archives are local or remote). So:

* **data interfaces**
* **VO services**
* **work-flows / pipelines**
* **data curation and preservation (DOI)**

The archive front end is a Science Gateway!

Data interface / VO Services

Web interface statistics:					
		LBTO	LBTB	INAF	
				Web	Vo
last week:	n of access:	66	0	197	606
	n of downlaod:	6	0	156	148
last month:	n of access:	1264	36	1272	2414
	n of downlaod:	268	13	1189	1175
Total:	n of access:	21250	4123	30611	78876
	n of downlaod:	5052	1323	25750	24940



IA2 - Vo Tools:

- ✓VODance – VO compliant data publication tool
- ✓Powered IA2TAP – VO compliant catalogs publication tool

But an archive is not only that...
You want more from your data!

Archive means having raw, science ready products (calibrated data, catalogs and so on..) possibly VO compliant and/or accessible by VO services (it doesn't mater if the archives are local or remote). So:

* **data interfaces**
* **VO services**
* **work-flows / pipelines**
* **data curation and preservation (DOI)**

The archive front end is a Science Gateway!

GAPS example:

GAPS, *Global Architecture of Planetary Systems*, is a long-term progra[illegible] comprehensive characterization of the architectural properties [illegible]netary systems as a function of the hosts' characteristics [illegible] metallicity, environment). Based on data acquired at HARPN @ TN[illegible]

- 50 INAF people + 15 international collaborators

GAPS Request:

1. Strong **interaction** with HARP[illegible] private data;

2. Customizable **re-p[illegible]** GAPS data with appropriate spectral line mask and option/s;

3. Per[illegible]**eries** on additional meta-data content;

[illegible] **repository** where **to access, synchronize, share and search** for interesting data;

5. A flexible and collaborative **tool to manage additional info** about the project and the observations.

Interaction between archive ad pipeline by Workflow management system (YABI)

But an archive is not only that...
You want more from your data!

Archive means having raw, science ready products (calibrated data, catalogs and so on..) possibly VO compliant and/or accessible by VO services (it doesn't mater if the archives are local or remote). So:

* **data interfaces**
* **VO services**
* **work-flows / pipelines**
* **data curation and preservation (DOI)**

The archive front end is a Science Gateway!

Data curation & preservation + Intellectual proprieties

Hot topics (H2020):

- Data Curation & Preservation;
- Open Access
- Open Source
- Intellectual proprieties

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.

Conclusions:

IA2 can actually provide:

- Archiving system (raw and calibrated) with NADIR;
- Data curation;
- Graphical User Interfaces;
- VO compliant data and catalogs publication;
- Work-flow system for pipelines management
- Data sharing area;

What IA2 will provide in the near future:

- New Graphical User Interfaces;
- SSO technology;
- DOI system adoption