Data Management and Archiving in Astronomy: Status and Challenges for the Future

Fabio Pasian

Istituto Nazionale di Astrofisica – O.A.Trieste

Euclid Consortium Science Ground Segment Manager

former Chair (2007-2010), International Virtual Observatory Alliance

ICTP, Trieste, 5 September 2013

FOREWORD

From this meeting’s Web page:

«Due to the increasing usage of computer simulations, the management of large amounts of scientific data has become a challenge.»

From the ESO Web site:

«The current total archive holding is about 65 TB, with an input rate of about 15 TB per year. This will soon drastically increase by a factor of 10 or so as the Visible and Infrared Survey Telescope for Astronomy (VISTA) with its near infrared camera will alone produce about 150 TB of data each year.»

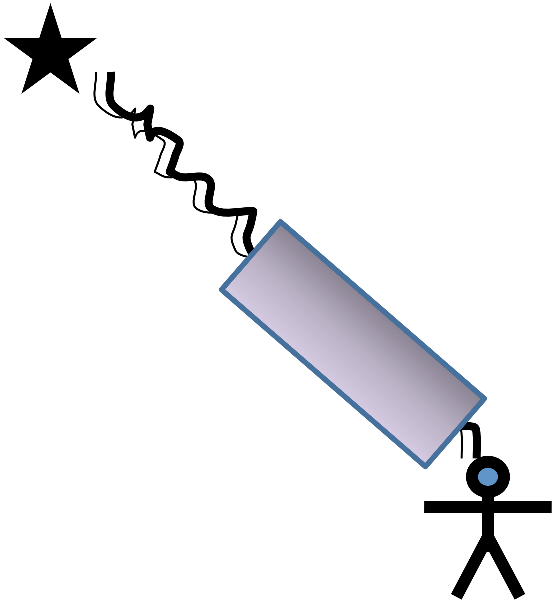
From the «SKA: the ultimate big data challenge» paper:

«The challenge? How to transport, store and process the estimated 14 exabytes of data the antennas will gather every day.»

SUMMARY

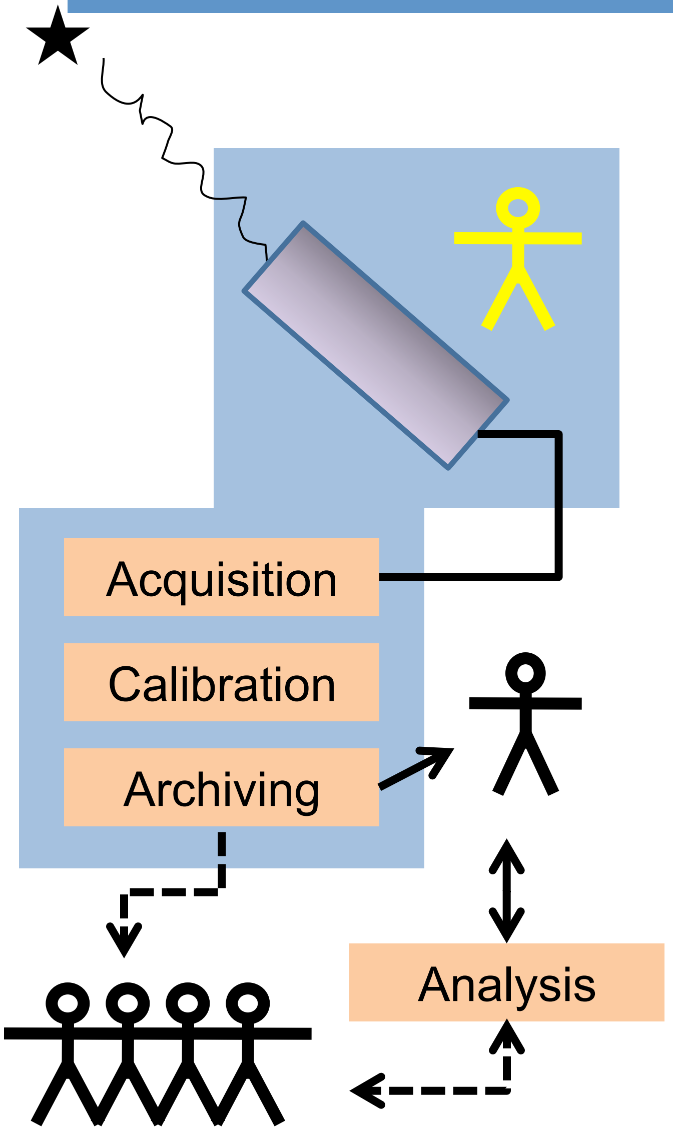
- Basics on astronomy
- Data sharing and standards
- The Virtual Observatory
- The big surveys
  - An example: the Euclid project
- … in Trieste

BASICS ON ASTRONOMY (I)



- Astronomy is an observational science - no repeatability
- Measures I(λ), or I(ν), or I(E)
- We actually observe I'(λ) = I(λ) * T(x,λ)
- Most phenomena are in fact variable
- We actually observe I"(λ,t) = I(λ,t)*T(x,λ,t)
- Every single observation must be kept ⇒ preservation!

BASICS ON ASTRONOMY (II)



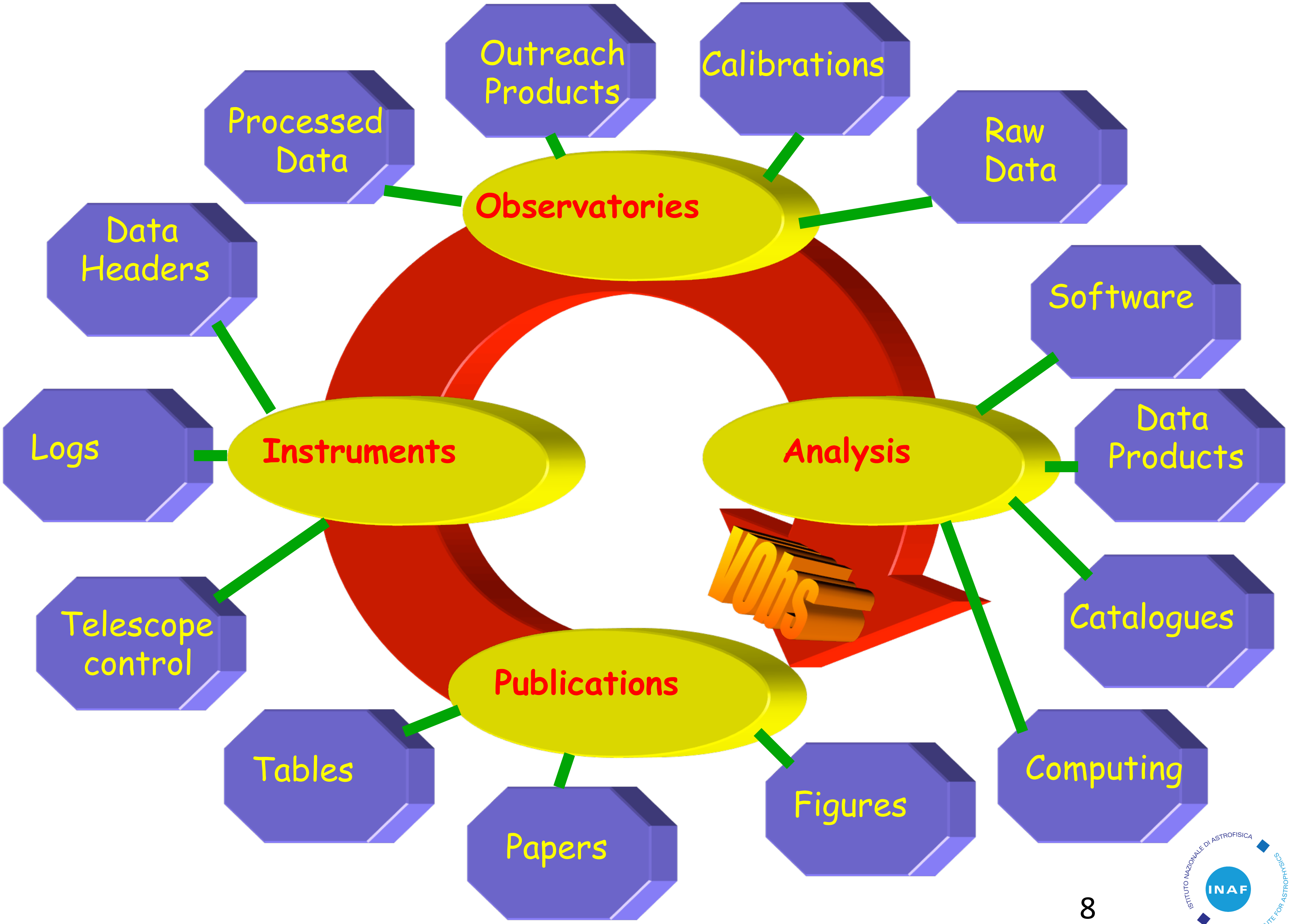
- Service observing due to:
  - complexity of instrumentation
  - optimisation of observational conditions
- Data handling and basic processing (calibration) c/o observatory
- Owner of observation gets data from archive
- After some time, data become public
- Since 1977, there is a standard format for data files (FITS)

FITS

- Flexible Image Transportation System
- Born in 1977 for images
- Later extended for N-dimensional data arrays, ASCII and binary tables, event lists, ...
- 3 continental committees, IAU WG, permanent FITS office (in GSFC/NASA) maintaining standards and sw
- Standard format
  - ASCII header containing keywords with data description
  - standard record size: 28800 bytes
  - readers available in any astronomical application
- (now used by Vatican Museum for digitised books)

NEED FOR ARCHIVES

- Monitor time variability of phenomena (digitization of photographic plates)
- Need to reprocess raw data given better knowledge of instrumental effects
- Compare phenomena in different bands (multi-λ astronomy)
- Increase return for investment (data re-use, educational, outreach, ...)
- Statistical analysis / mining of large quantities of data
- Cope with data avalanche



Observatories

Processed Data

Outreach Products

Calibrations

Raw Data

Analysis

Software

Data Products

Catalogues

Computing

Publications

Figures

Papers

Tables

Instruments

Telescope control

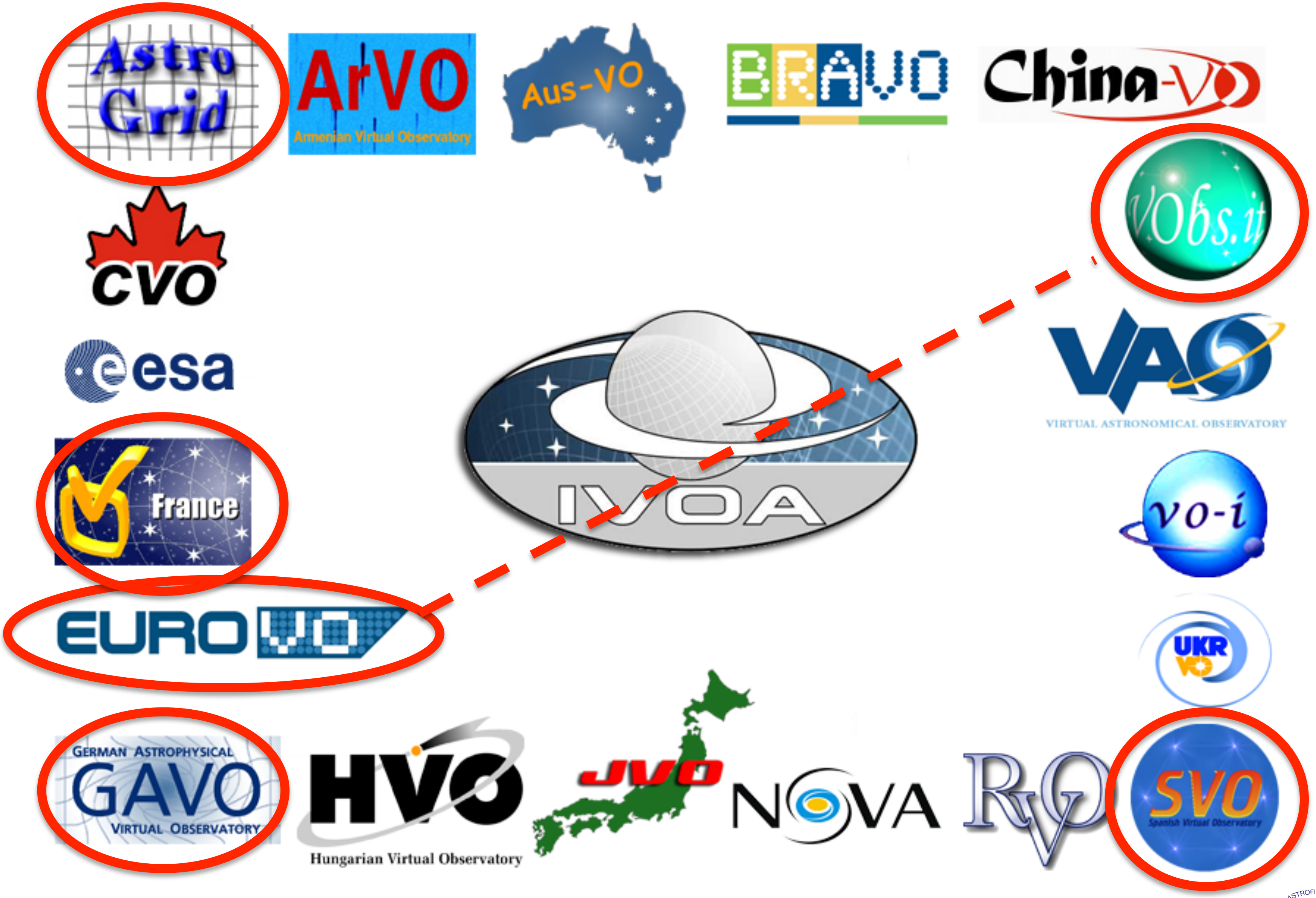
Logs

Data Headers

THE IVOA

- **Mission:** *"To facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory"*
- Works by telecons, "Wiki" pages, and bi-annual meetings (last one in Heidelberg [May 2013], next in Waikoloa [Sept 2013])
- **Needs:** standardization of data/metadata/sw, data interoperability methods, and list of available services (provided by projects)
- Structure:
  - ✓ IVOA Executive Board includes representatives from all VObs projects
  - ✓ Working and Interest Groups (including Theory)

http://www.ivoa.net



Euro-VO : 5 EU/FP funded projects (2004-2014)

ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD) – COMMENTS ON THE VOBS

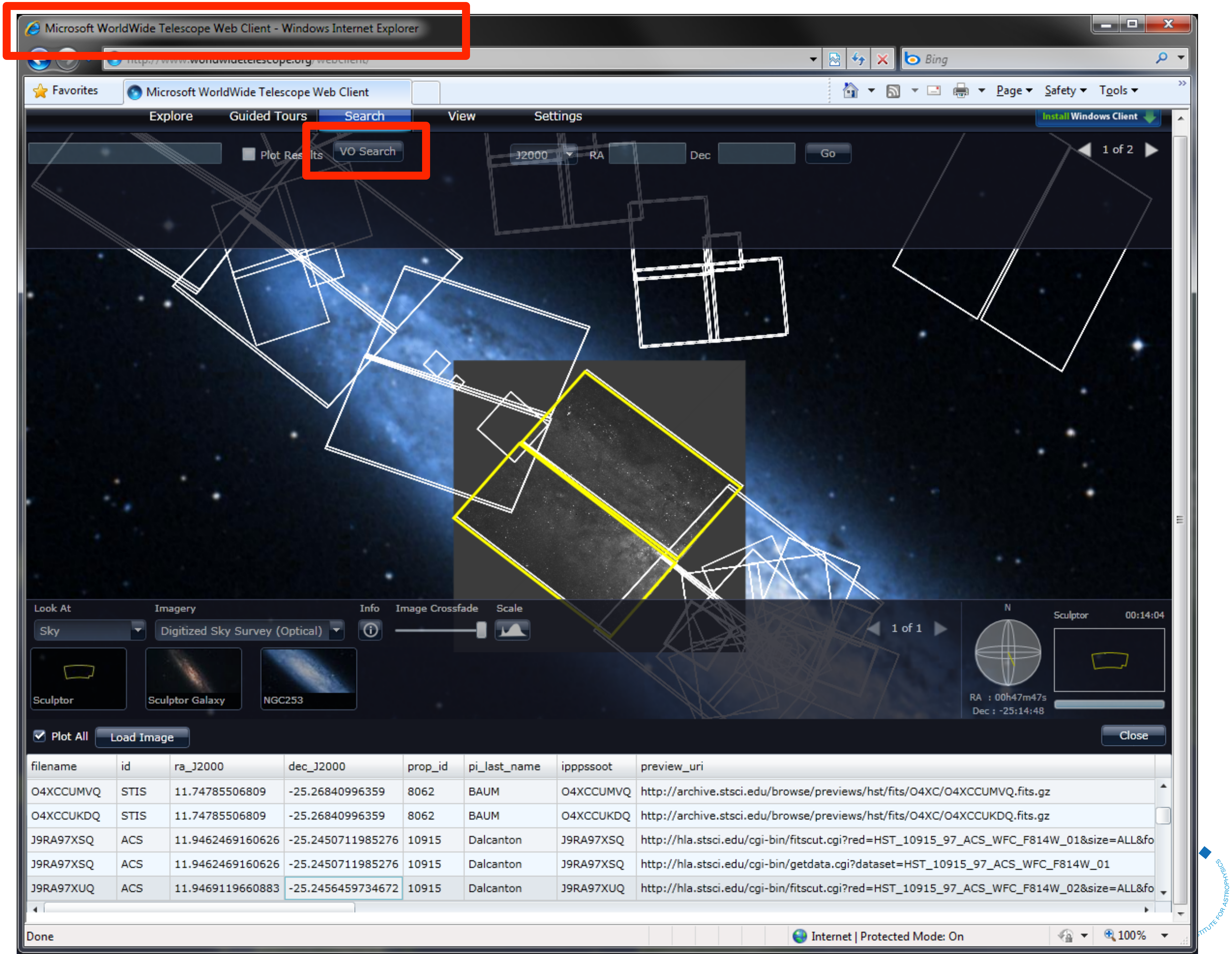
- Findings
  - The Virtual Observatory concept is a bold community-led response to the challenges the astronomical community faces in data management and storage. Impressive progress has been made and the momentum of the International Virtual Observatory Alliance will ensure sustained progress, provided the agency level support and funding is available.
- Recommendations
  - New projects and facilities must take the data management, storage, maintenance, and dissemination into account at the earliest planning stages, consulting potential users in the process. Agencies should recognise that this is an important long term issue and should co-ordinate plans, provide adequate funding on a long-term basis, and support development and maintenance of the needed infrastructure. *Agencies should encourage the broadening of the existing VObs collaboration into a fully representative global activity*.

ESFRI and ASTRONET statements

- ESFRI (multi-disciplinary)
  - focus on networking, capability & throughput computing, grid architectures, software, data management and curation

  → Research Data Alliance

- ASTRONET (Astronomy & Astrophysics)
  - recognised as must-haves to tackle the challenges of the future (priority in resources):
    - computing (capacity AND capability)
    - theory & simulations
    - virtual observatory
    - laboratories



Look At: Sky

Imagery: Digitized Sky Survey (Optical)

Sculptor 00:14:04

RA : 00h47m47s

Dec : -25:14:48

☑ Plot All

filename	id	ra_J2000	dec_J2000	prop_id	pi_last_name	ipppssoot	preview_uri
O4XCCUMVQ	STIS	11.74785506809	-25.26840996359	8062	BAUM	O4XCCUMVQ	http://archive.stsci.edu/browse/previews/hst/fits/O4XC/O4XCCUMVQ.fits.gz
O4XCCUKDQ	STIS	11.74785506809	-25.26840996359	8062	BAUM	O4XCCUKDQ	http://archive.stsci.edu/browse/previews/hst/fits/O4XC/O4XCCUKDQ.fits.gz
J9RA97XSQ	ACS	11.9462469160626	-25.2450711985276	10915	Dalcanton	J9RA97XSQ	http://hla.stsci.edu/cgi-bin/fitscut.cgi?red=HST_10915_97_ACS_WFC_F814W_01&size=ALL&fo
J9RA97XSQ	ACS	11.9462469160626	-25.2450711985276	10915	Dalcanton	J9RA97XSQ	http://hla.stsci.edu/cgi-bin/getdata.cgi?dataset=HST_10915_97_ACS_WFC_F814W_01
J9RA97XUQ	ACS	11.9469119660883	-25.2456459734672	10915	Dalcanton	J9RA97XUQ	http://hla.stsci.edu/cgi-bin/fitscut.cgi?red=HST_10915_97_ACS_WFC_F814W_02&size=ALL&fo

BIG SURVEYS

- Digitized Sky Survey (DSS): a digital version of several photographic atlases of the sky
- Sloan Digital Sky Survey or SDSS: major multi-filter optical imaging and spectroscopic redshift survey
- PanSTARRS, DES, LSST, …
- ROSAT All Sky Survey: 1378 distinct fields in X-rays
- Planck: whole sky in 9 λ bands from radio to infrared

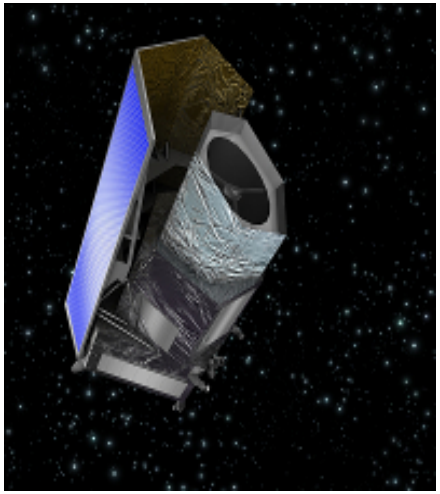
→ Euclid: whole extragalactic sky (15000 degrees^2) in visible and IR imaging, IR spectra

- ALL DATA ARE INTEROPERABLE USING VOBS

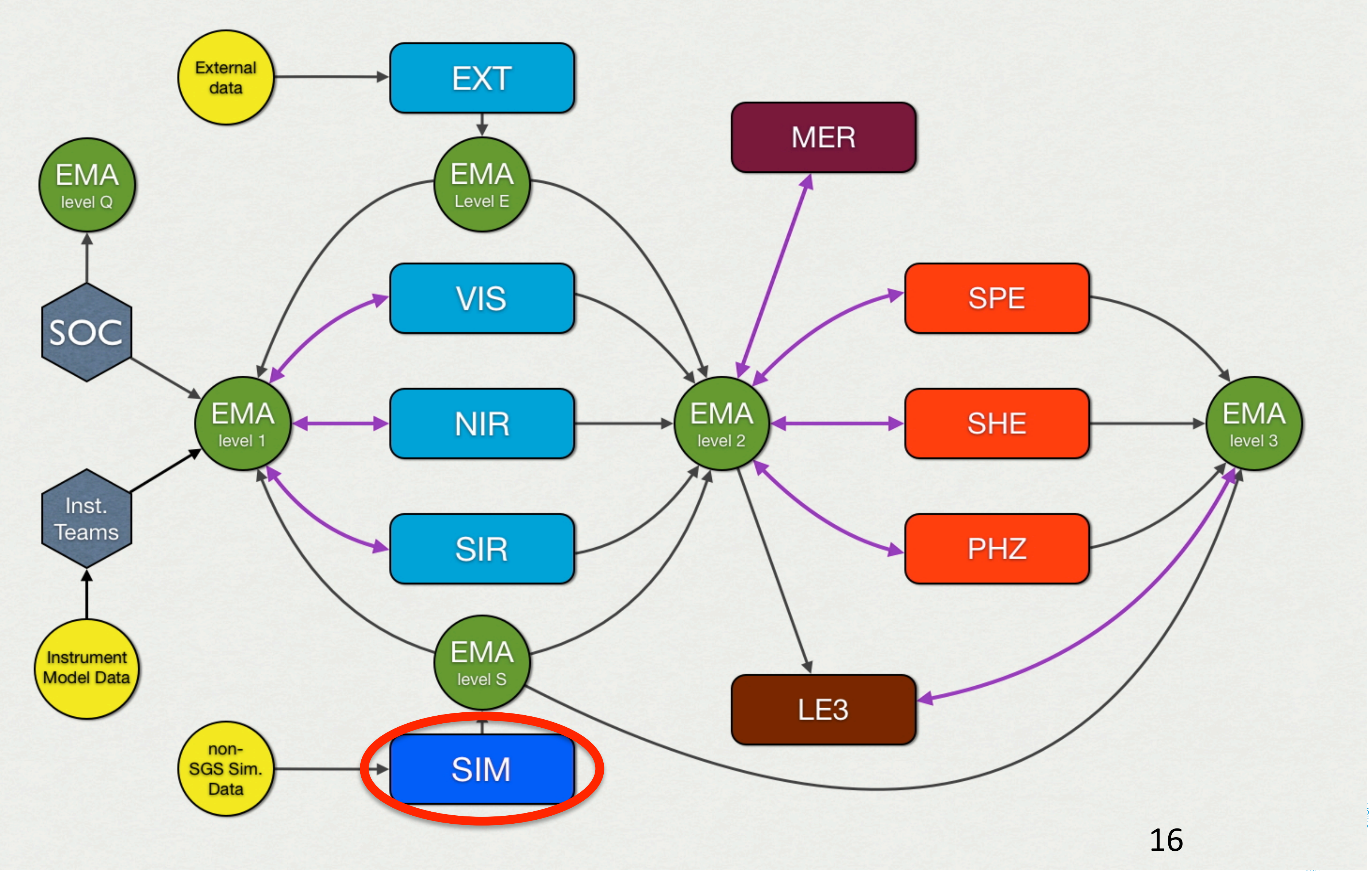
THE EUCLID MISSION

How did the Universe originate and what is it made of?

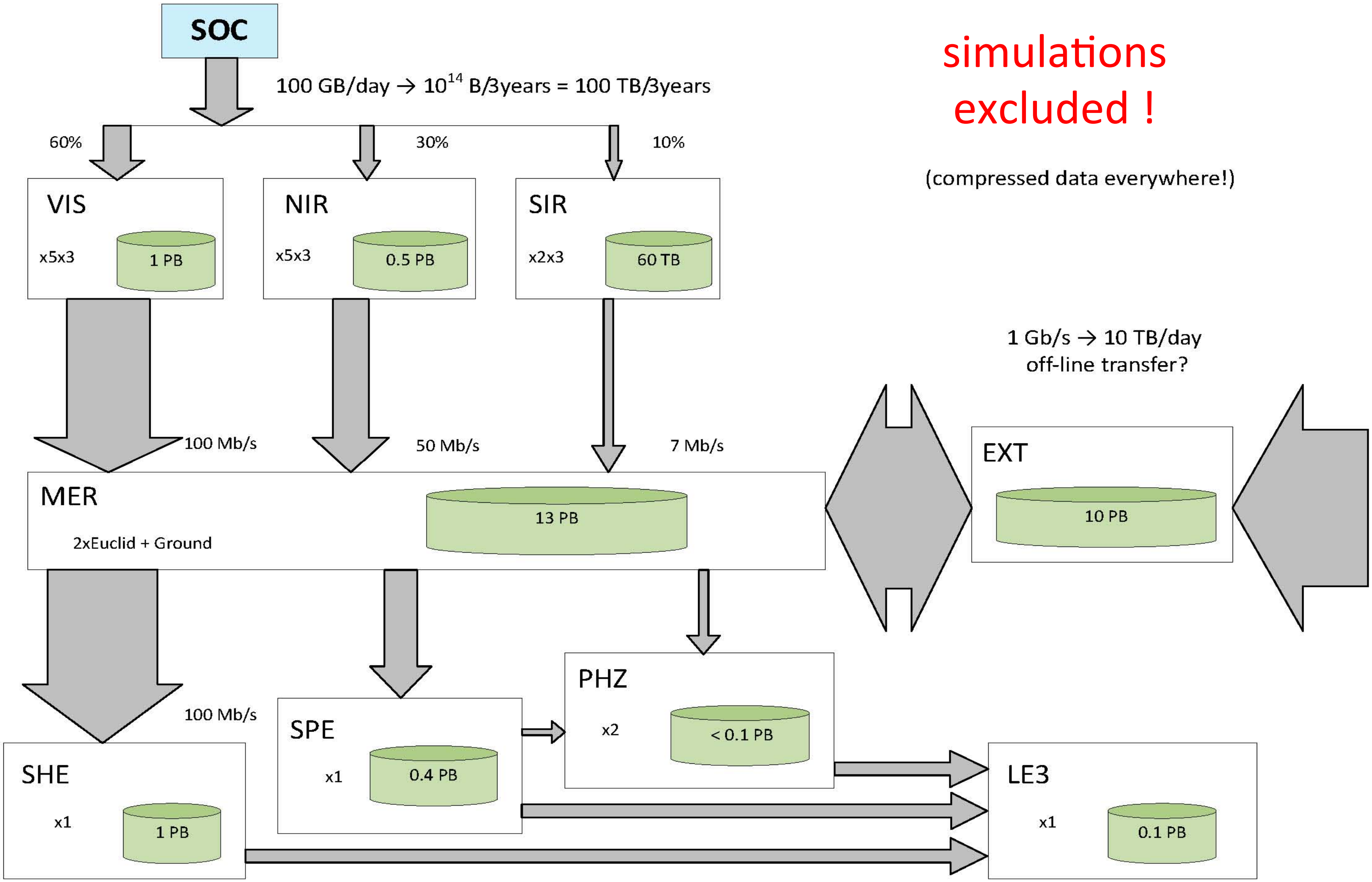
- Goal: To understand the nature of dark energy and dark matter by accurate measurement of the accelerated expansion of the Universe
- Targets are galaxies and clusters of galaxies out to z~2, in a wide extragalactic survey covering 15 000 deg^2, plus a deep survey covering an area of 40 deg^2
- Two instruments (NIR imager and spectrograph + visible imager) + need for external data
- Approved by the European Space Agency (Oct 2011)
- Date of launch: March 2020



EUCLID DATA PROCESSING FUNCTIONS



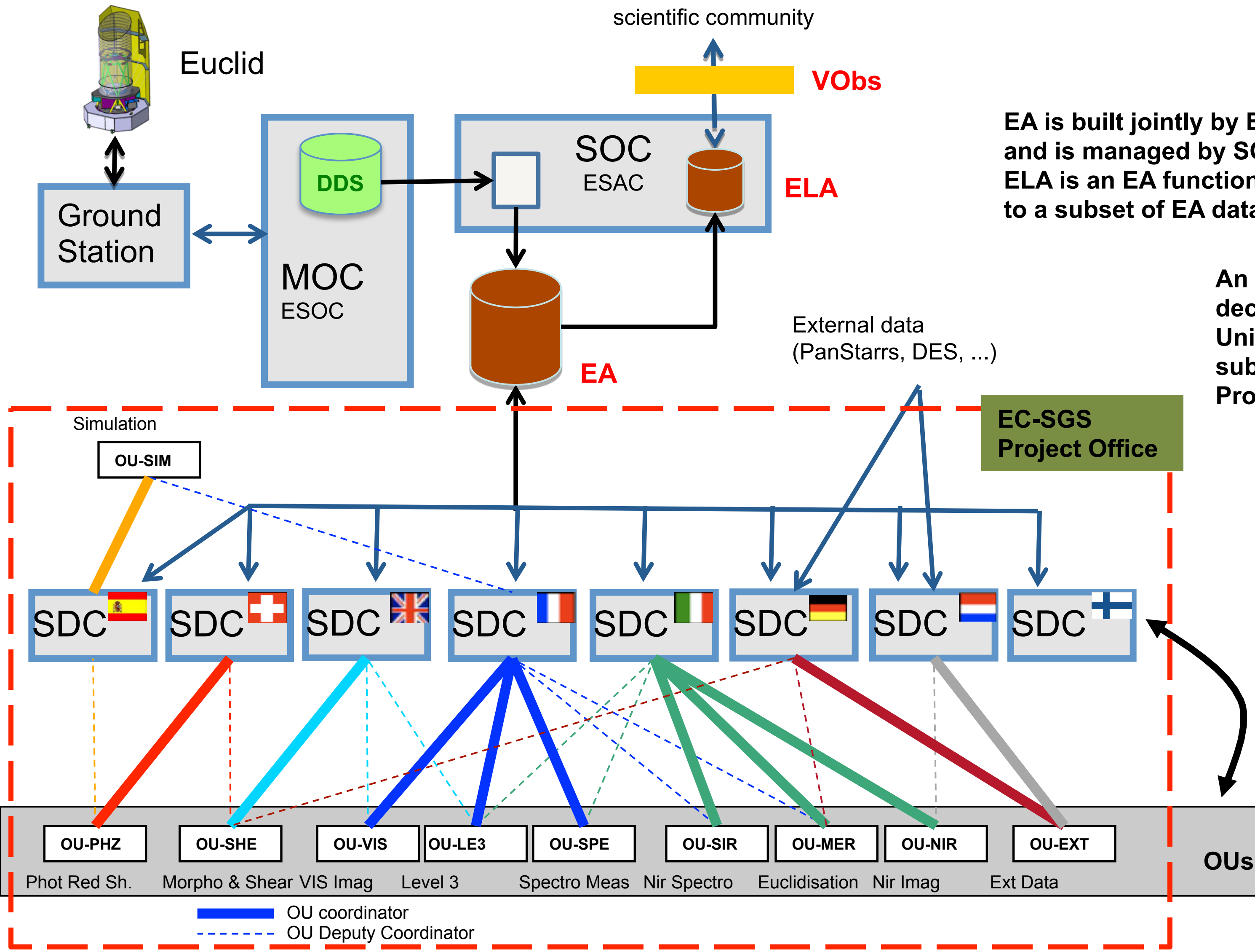
EUCLID DATA FLOW



simulations excluded !

(compressed data everywhere!)

EUCLID GROUND SEGMENT: OVERALL VIEW



EA is built jointly by EC and SOC, and is managed by SOC
ELA is an EA function, allowing public access to a subset of EA data

An organization based on the decomposition in Organization Units (OU), corresponding to a subset of overall EUCLID Data Processing.

Each OU produces algorithms which are integrated and executed in an SDC (Science Data Center)

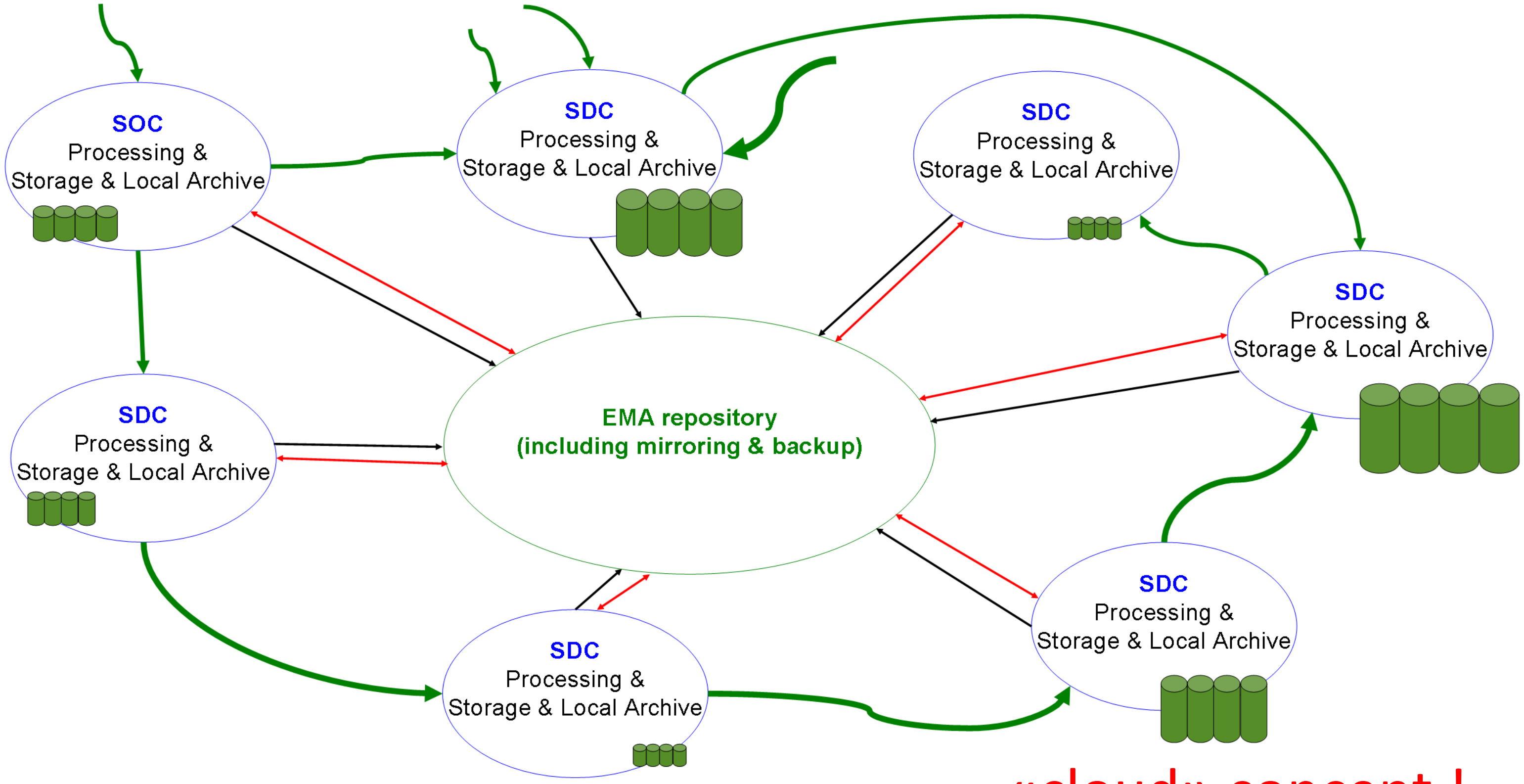
The SGS System Team provides support and tools for the whole of the SGS (SOC + EC-SGS)

OUs are transnational

EUCLID ARCHIVE SYSTEM

- For internal use within the Euclid Consortium:
  - Centralised metadata repository (at ESA) based on DBMS
  - Distributed data files (images, spectra, catalogues)
  - At the core of the (data-centric) Euclid processing:
    - Processing «pipelines» work on fractions of the sky stored on local computing centre
    - Data replicated at some other centre as necessary
- Public data are replicated at the European Space Astronomy Centre of ESA (ESAC) and from there are made accessible to the external community through VObs-compliant interfaces.

EUCLID DATA + PROCESSING LOGICAL ARCHITECTURE



Data Product (image) transfer between SDC's

Metadata update

Query/Metadata

«cloud» concept !

full pipeline is run on sky tiles

virtualisation – Hadoop?

IN ITALY (… actually, led by TRIESTE)

- Coordination of INAF ICT Unit
- INAF centre for Astronomical Archives (IA2) and VObs.it (~ 150 TB)
- Planck- LFI Data Processing Centre (~ 250 TB)
- In Euclid:
  - management of Science Ground Segment (Project Office)
  - design, development and operations of SDC-Italy
  - development of NIR, SIR, MER, SPE, LE3, SIM processing
  - coordination of the infrastructure needed for simulations
- Participation in MIUR-approved DHTCS-IT project (Distributed High-Throughput Computing System) → cloud for research
- Submitted projects to MIUR («premiali»):
  - AstroCloud (INAF-led, with UniNA and SISSA)
  - participation in PIDES (multi-disciplinary data preservation)
- Participation in RDA and CINECA «Big Data» initiatives

CONCLUSIONS

- Increasing detector performance + astronomical all-sky surveys: from Big Data to **Bigger Data** (VObs-compliant)
- Theory data and modelling add up to the data volume – there is no limit to the data that can be produced, the physics to model is already there – (VObs-compliant)
- Individual projects may be able to cope through a cloud-like approach
- «What can we afford to throw away?»
- Long tradition in data sharing (VObs) → standards (they are a blessing, but also a constraint)
- There is political rather then technical issue: how to perform massive data analysis within the VObs?

THANK YOU
FOR YOUR ATTENTION !

fabio.pasian@inaf.it
riccardo.smareglia@inaf.it