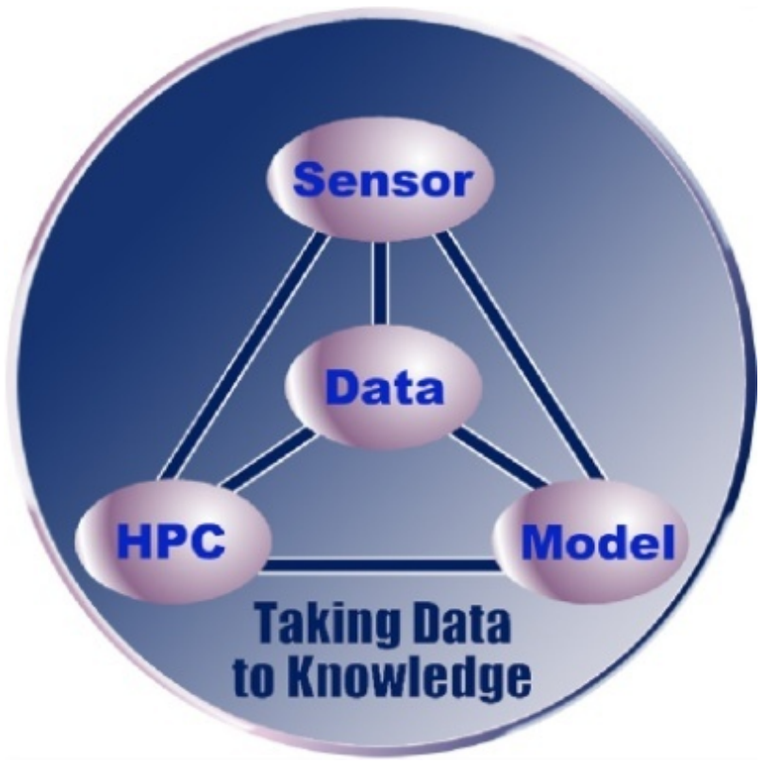
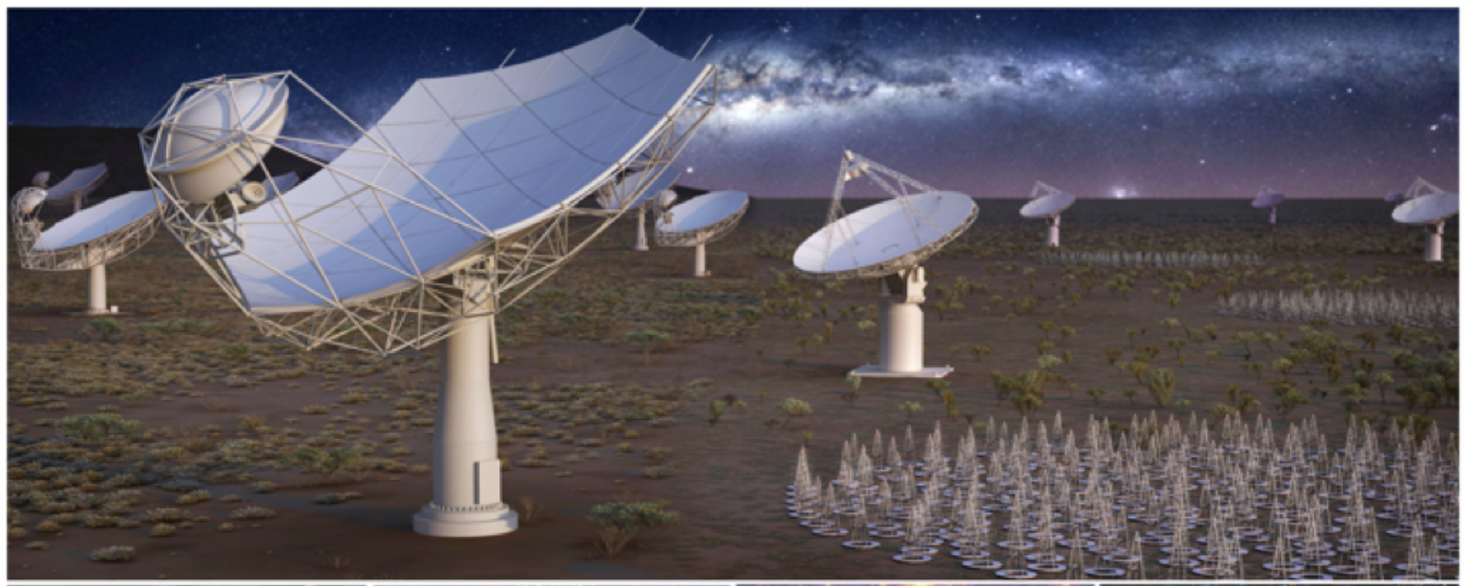
France within ESRC, SRCs and SKA

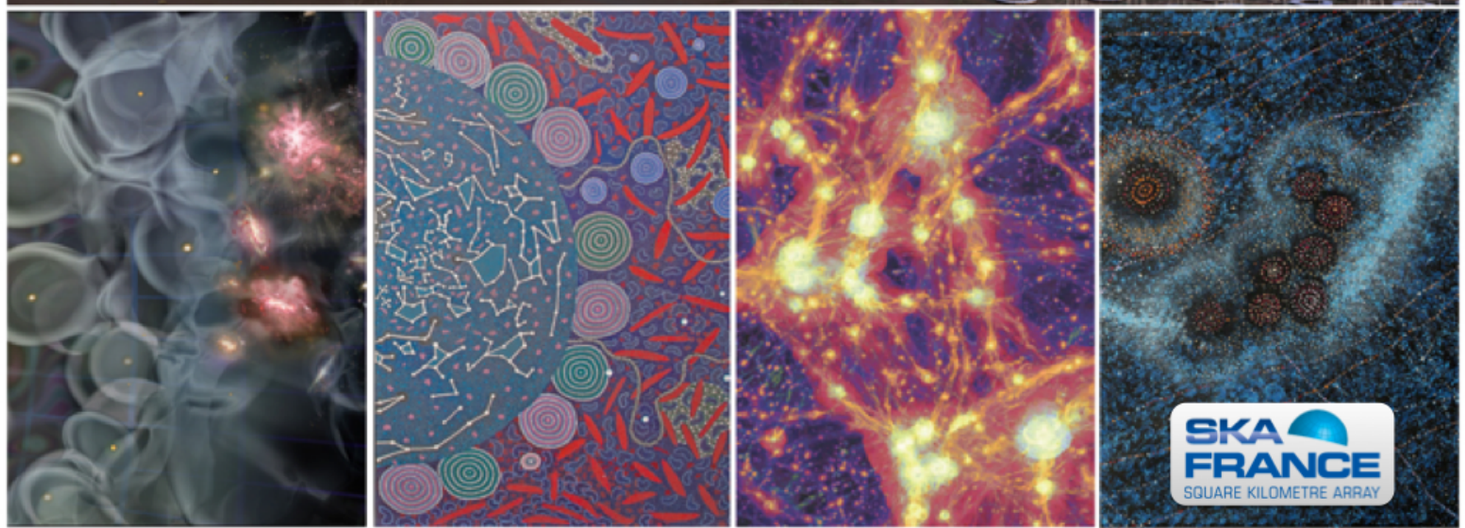
Jean-Pierre Vilotte

Scientific Deputy (Data & HPC)

Institut des Sciences de l'Univers (CNRS-INSU)







SKA: France landmarks and roadmap

July 2016: SKA-France academic coordination
- leaded by CNRS-INSU

October 2017: SKA-France white paper publication
- 176 contributors from 40 research groups
- 6 private entreprise

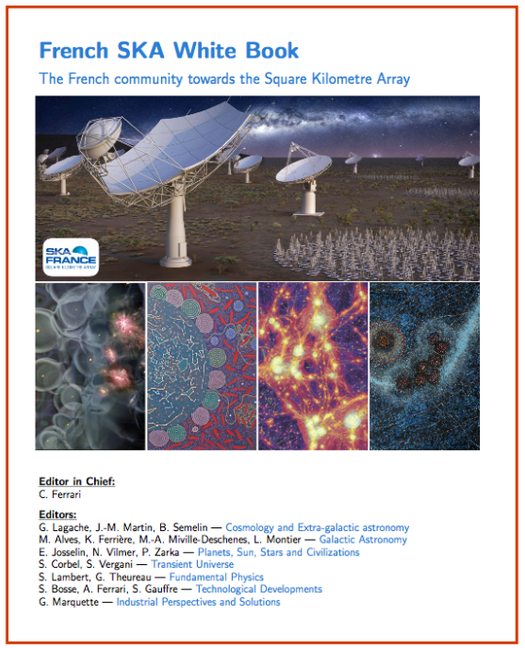
February 2018: Maison SKA-France (MSF)
- real equilibrated PPP: research organisations, industry partners (expanding, e.g. INRIA, Engie, Total, EDF, Ariane group)
- science and technology thinking: working groups, conferences, monthly bulletin, strategic documents
- forum to develop fundamental research and R&D projects
- new business model for large research infrastructures (TGIR)

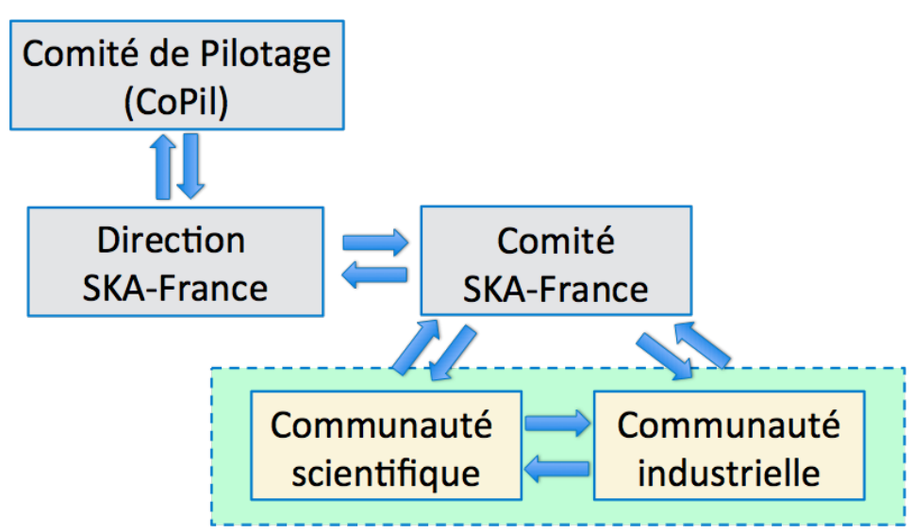
Mai 2018: SKA in the TGIR French roadmap (MESRI)

July 2018: CNRS as special member of SKAO (3 years)
- involving the MSF

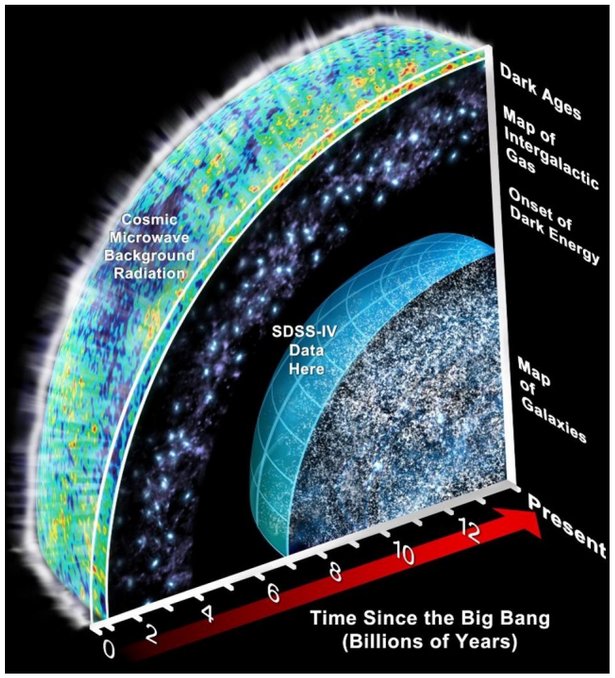
2019-2020: Revision of the French TGIR roadmap
- CNRS & MSF coordinates preparation of French insertion within SKAO, and SRCs and ESRC infrastructures

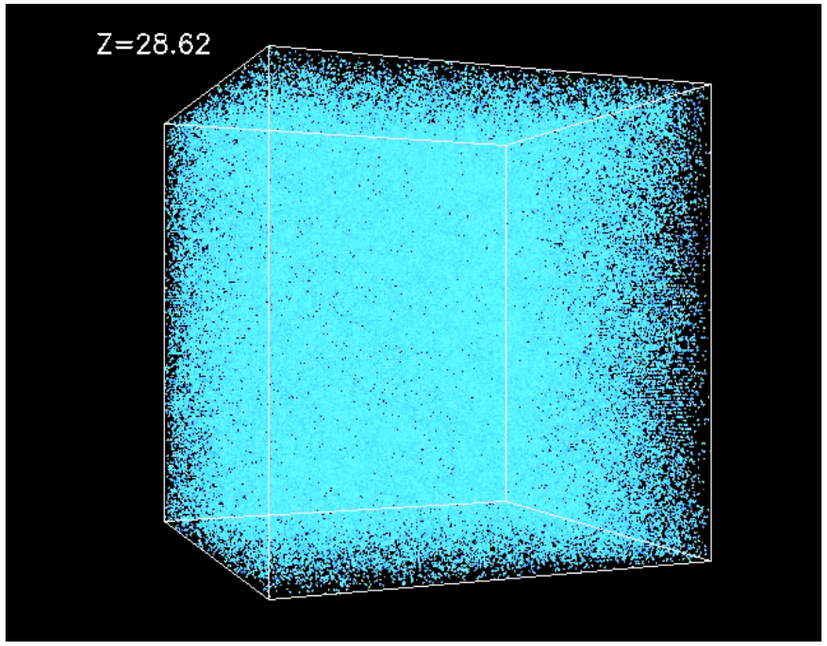
2020-2021: National decision to join SKA
- ***based on the evaluation of both SKAO and SRCs***





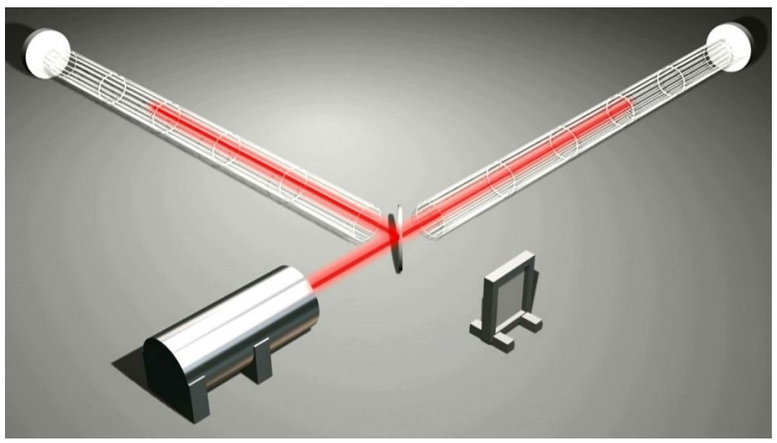
Astronomy and SKA

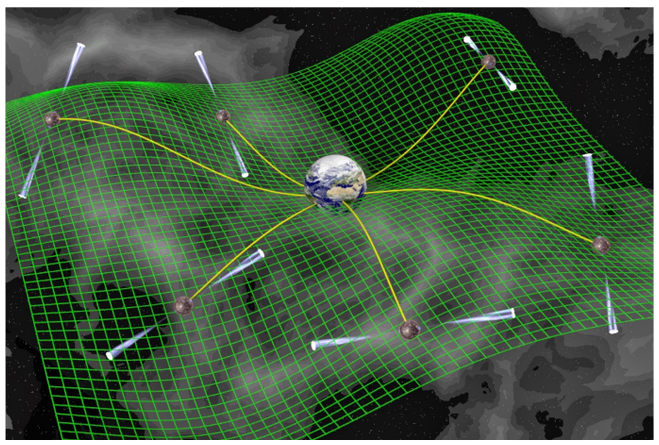




The era of big surveys

EUCLID: ~100 GB/day, 200TB over 5 yrs compressed detector data
LSST: ~160 MB/s, ~1.3 TB/night ,~ 30 PB over 5 yrs archived data
LOFAR: ~100 TB/night, ~6-10 PB/yr archived data
MDWA: ~16 GB/s, 6 PB/yr archived data
ASKAP: spectral line 80 PB/yr, Cont./HI sky 1-4 PB/yr
CTA: ~3-10 PB/yr archived data
SKA: ~0.1-3 EB/yr archived science-ready data





Cosmic dawn
(First stars & Galaxies)

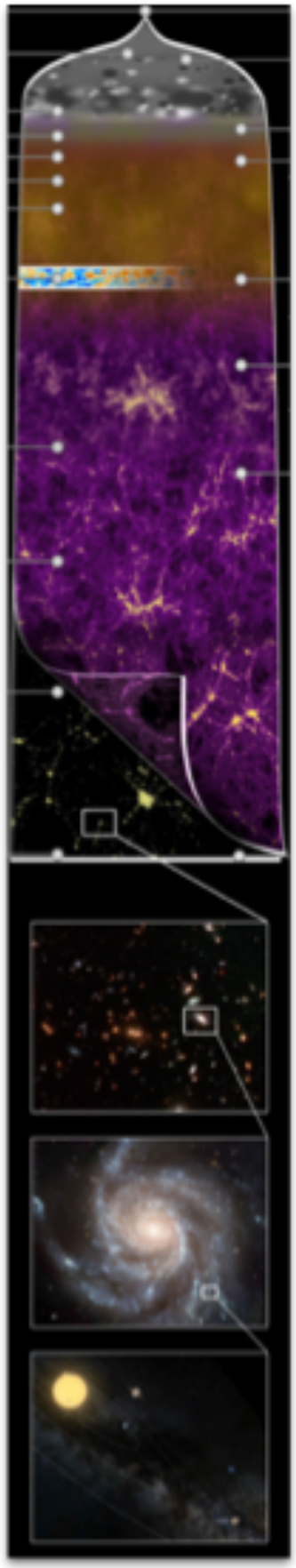
Cosmology
(Dark matter, Large-scale structures)

Galaxy evolution
(gas content & new stars)

Cosmic magnetism
(origin & evolution)

Fundamental physics
(gravitational waves & compact objects)

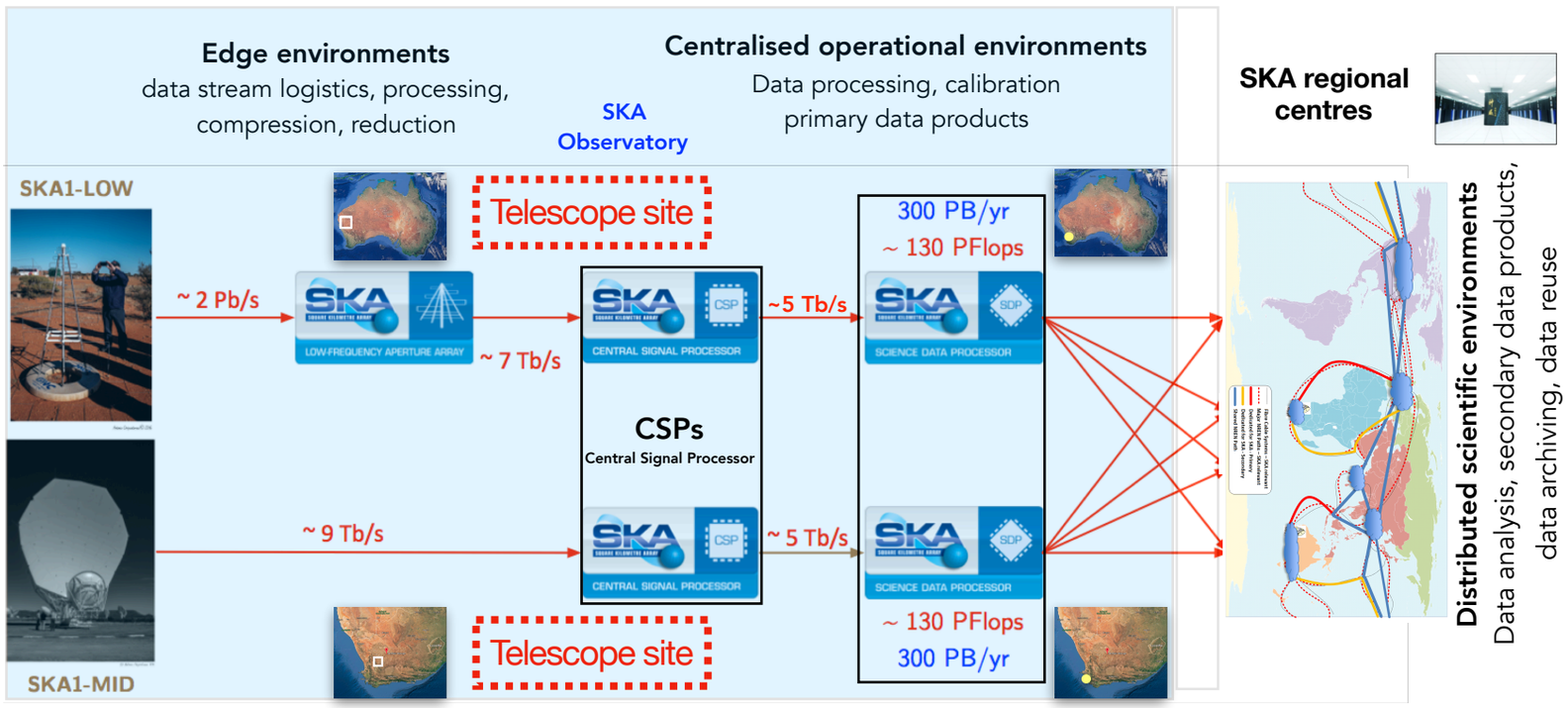
Cradle of life
(Planets, Molecules, SETI)



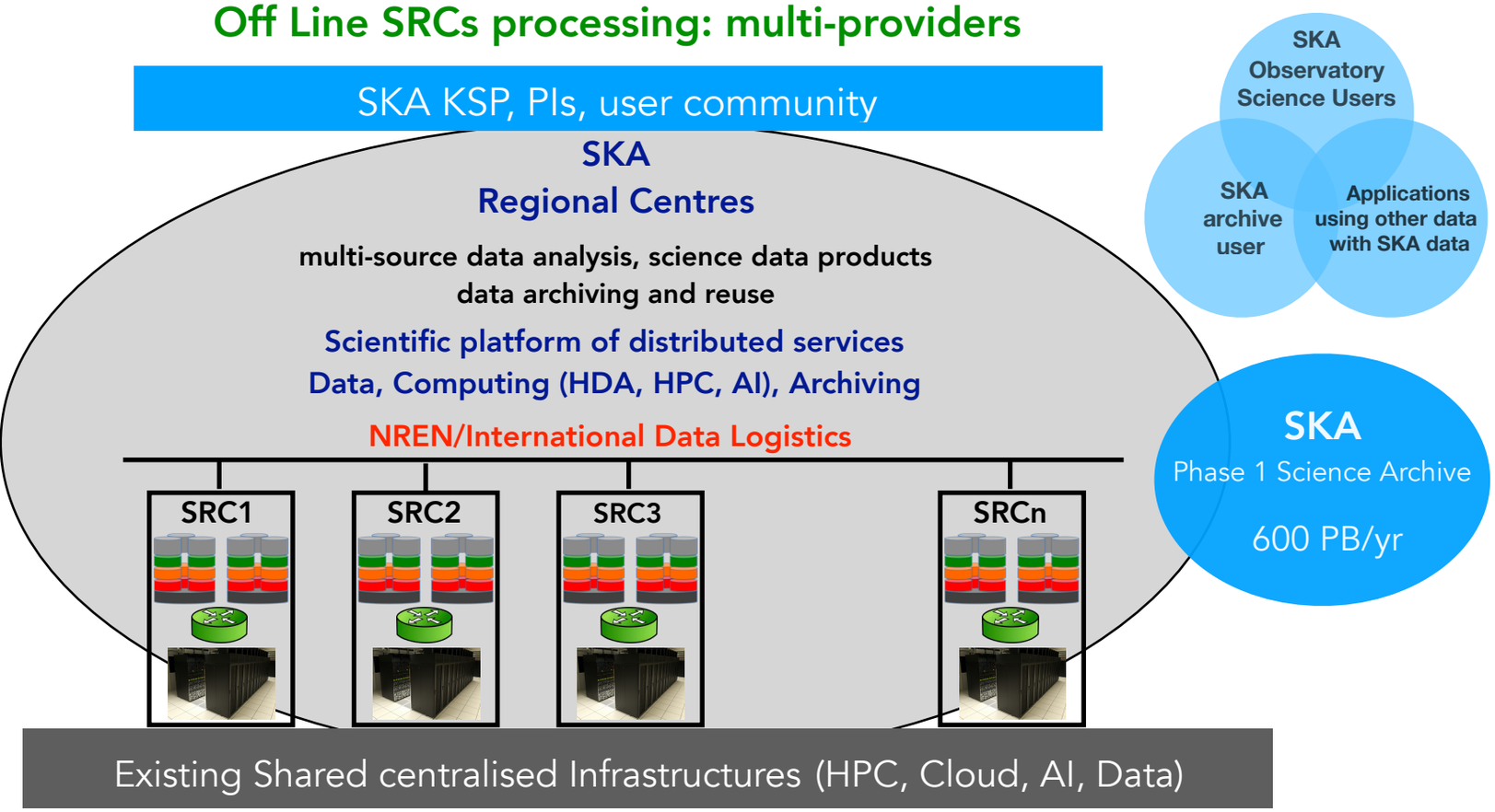
Modern sky surveys: ~10^{12} - 10^{18} bytes images; Catalog: ~10^{8}-10^{9} objects (stars, galaxies, etc.)

SKA: BigData community-driven strategy & path finder

On Line processing: high-rate data streams



Off Line SRCs processing: multi-providers



SKA observatory

From edge -> centralised infrastructures

- High-rate data stream logistics
- Stateful network services : caching/buffering
- Edge computing: numerical beam forming of signals ; removal of radio-frequency interference
- Data loss-compression and reduction
- Dynamic stream structures: observation dependent

Centralised HPC/HDA operational infrastructures

- Storage and computing capabilities/capacities
- High-rate data processing
- Complex HDA workflows (processing & calibration)

Primary data productS (events, images, cubes)

- Data models (standards, metadata, provenance)
- Archiving and dynamic distribution (data placement)

- > **Machine Learning moving to the edge**

SKA Regional Centres

New organisational, operational, business model

- Community-driven shaping strategy
- Co-designed (SKAO, providers, scientific users)

Scientific software platform

- Distributed services across shared infrastructures
- Multi providers (Cloud, HPC, Data), Federated AAI
- Application-dependent global ressource optimisation

Application workflows

- Diversity of complex workflows (HDA, HPC, AI)
- Data logistics all along in multi-provider context
- Workflow management and provenance system

Data archiving, curation and reuse

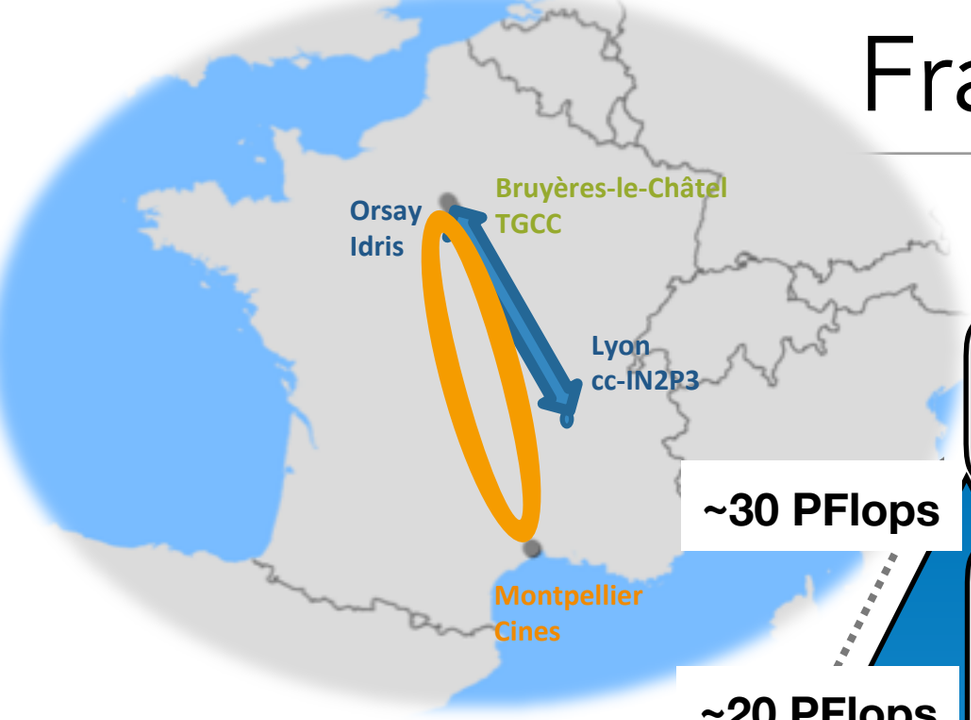
- Primary and secondary scientific data products
- FAIR multi-source data services (federated)

Scientific Users

- Key SKA Projects and PI granted observation projects
- Reuse of SKA data products: multi-messenger and multi-wavelength approaches

Shared with other communities: Earth & Space Observation, HEP

France SRC: Computing landscape



TGCC: renewed 2018
extension 2020

IDRIS: renewed 2019
extension 2021

CINES: renewal 2020
extension 2023

~30 PFlops

~20 PFlops

~1 PFlops

<500 TFlops

RENATER

European Governance

Tier-0 : TGCC@CEA (PRACE)

National Governance: HPC/HDA + AI

Tier-1 : IDRIS@CNRS, CINES@MESRI,TGCC@CEA

CC-IN2P3@CNRS (Tier-1 CERN)

National/Regional Governance

Tier-2 : **Meso-centres (HPC/HDA/AI)**

- MesoPSL (Paris)
- SIGAMM (Nice)

Regional/Local Governance

Tier-3 : **Computing and Data Analysis Platforms**

- Federation of Research
- Observatories in Science of Universe (OSU)
- Research Institutes

MESRI, CNRS, CEA, INRIA, Universities

CNRS

GENCI/CNRS Regions Universities

CNRS Universities

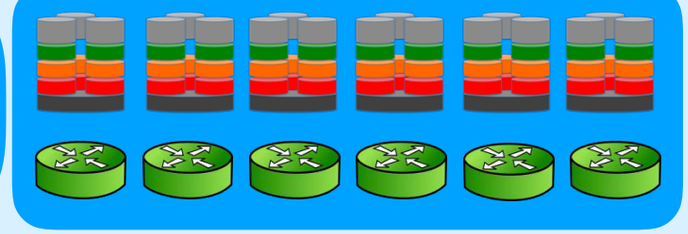
TGIR

France SRC: DC landscape

Euclid, LSST, CTA, SKA (?)

Tiers-1- DC1 National Hosting Compute & Data Centres

IDRIS@CNRS
CINES@MESRI
CC-IN2P3@CNRS



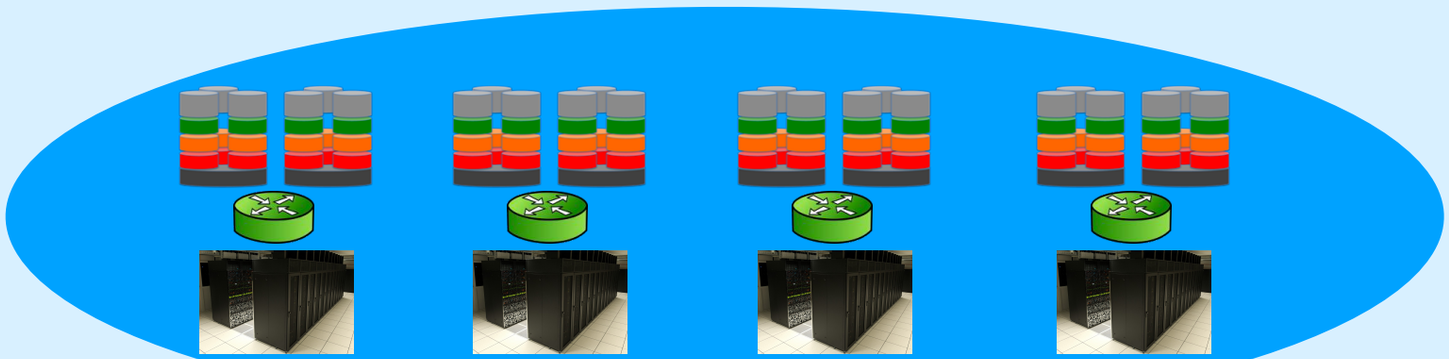
ESFRI/FR, TGIR, IR, Research Communities

RENATER
dedicated, secured,
high-bandwidth

Federated AAI

Tier-2 - DC2 Regional Hosting Compute & Data Centres

USN Data Centre
(LOFAR/NenuFAR)
Grenoble DC
(IRAM ARC node)
CDS/IVOA
(VO, catalogue)
...



GEANT

PRACE
EUDAT
EGI
GoFAIR
ESFRIs
...

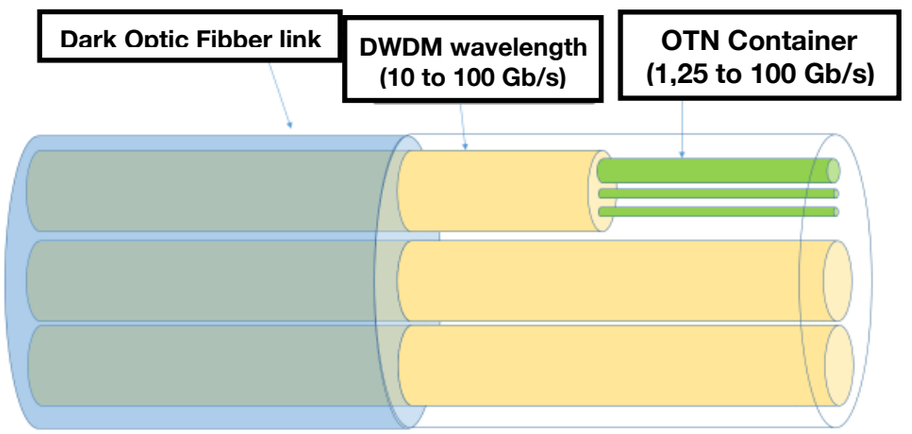
European Open Science Cloud

International e-Infrastructures

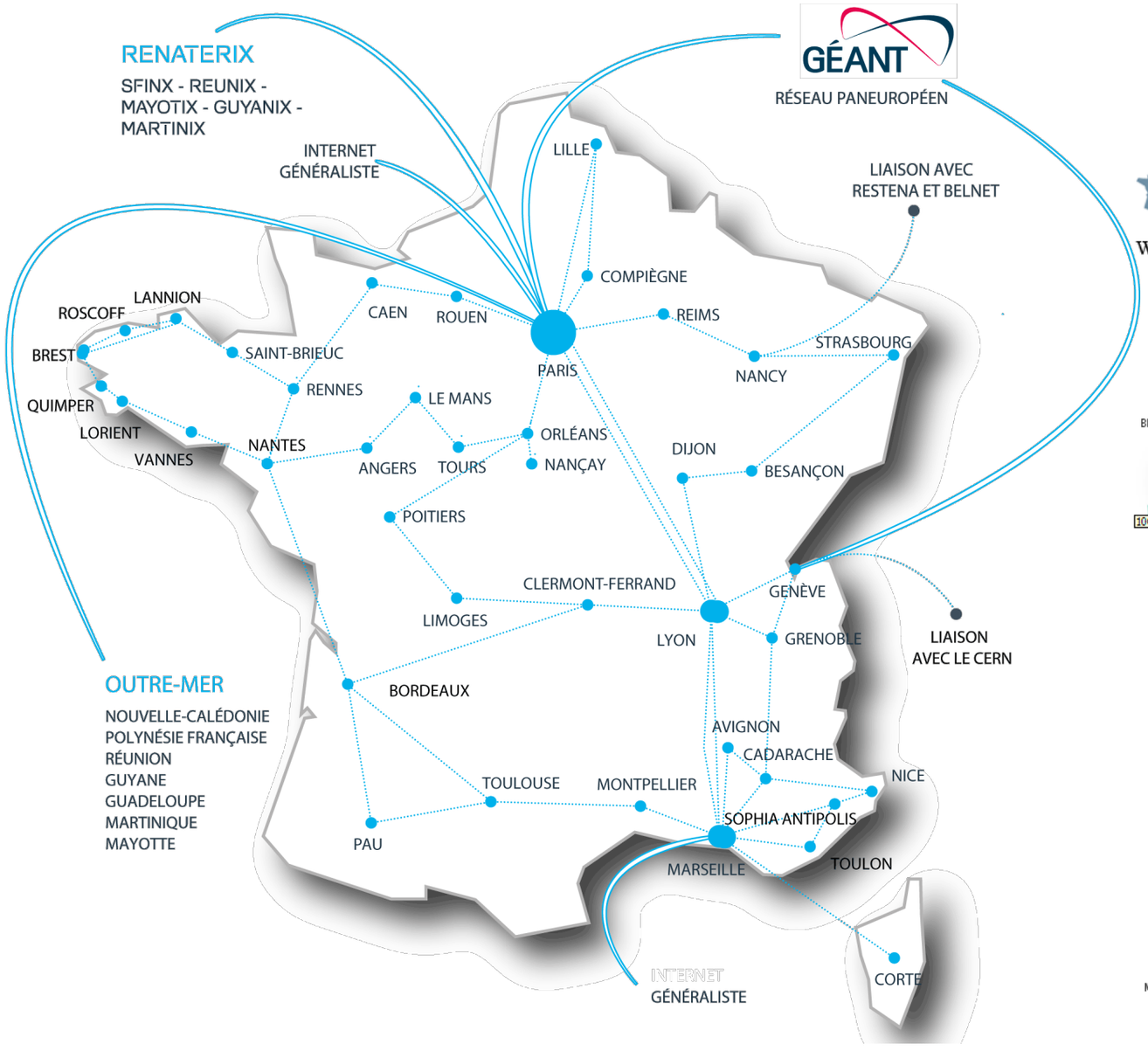
European Compute & Data Infrastructure

(SKA/SRCs, ALMA/ESO ...)

France SRC: Network landscape

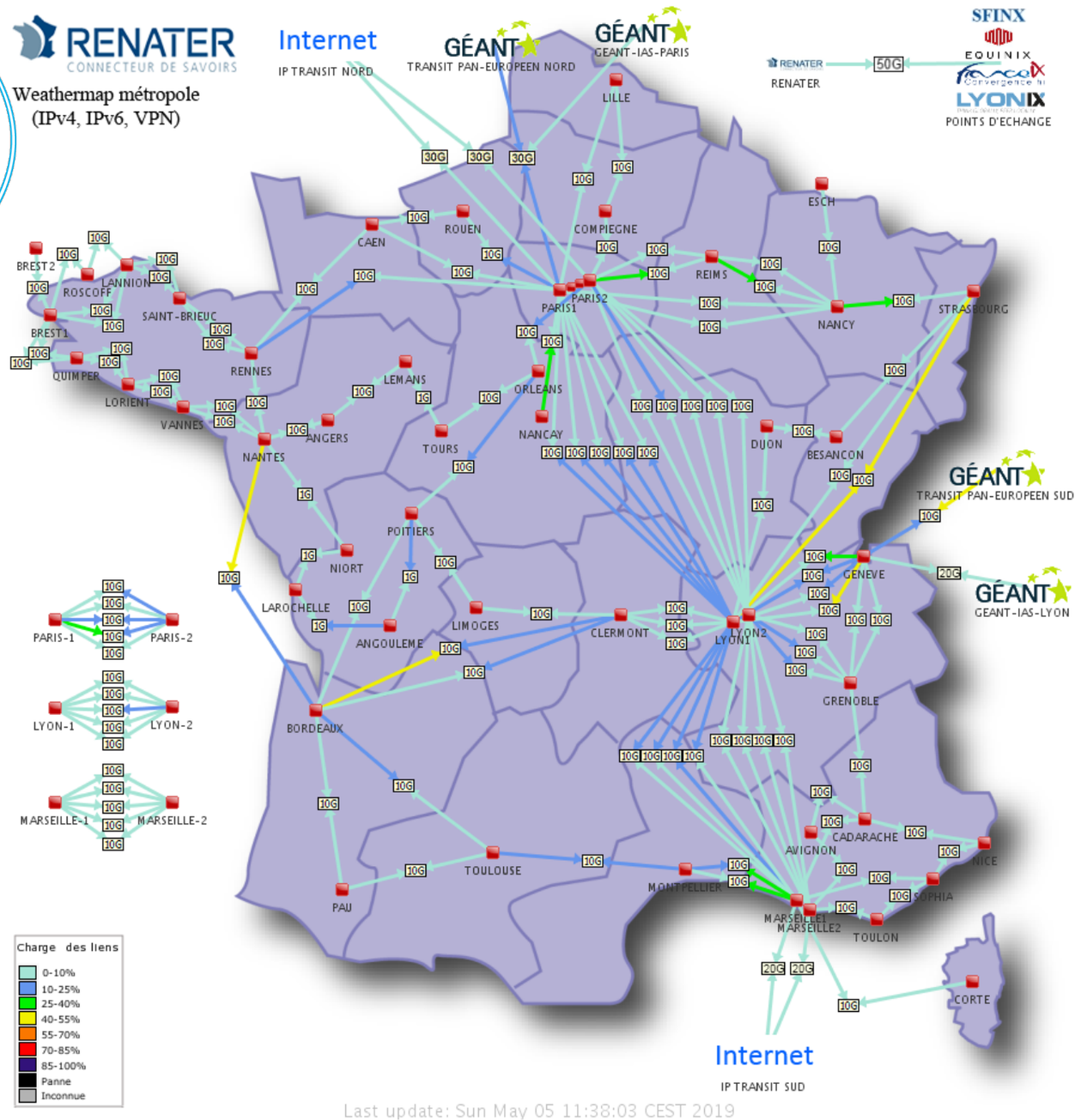


Physical links



- Dark fibbers support dedicated link capacities to projects (e.g. LHC-One, CCFR, NenuFAR…)
- On demand IP capacities: aggregated 10G link, 100G link on target …

Generalist IP services snapshot



Concerns and expectations

SRC governance, organisational and business models

- leverage regional & national policies & infrastructure investments
- funding agencies & cyber-infrastructure providers
- Science-driven optimisation of the distributed services and infrastructures capacities and capabilities
- agile and inclusive strategy

Include different SRC level of concern

- KSP and PI projects analysis of SKA (observatory) primary data products -> *coordinated with SKAO*
- Reuse of SKA data products (multi-messager, multi-wavelength) -> *community-driven shaping strategy* (SKA one component)

Develop a software platform model of distributed services

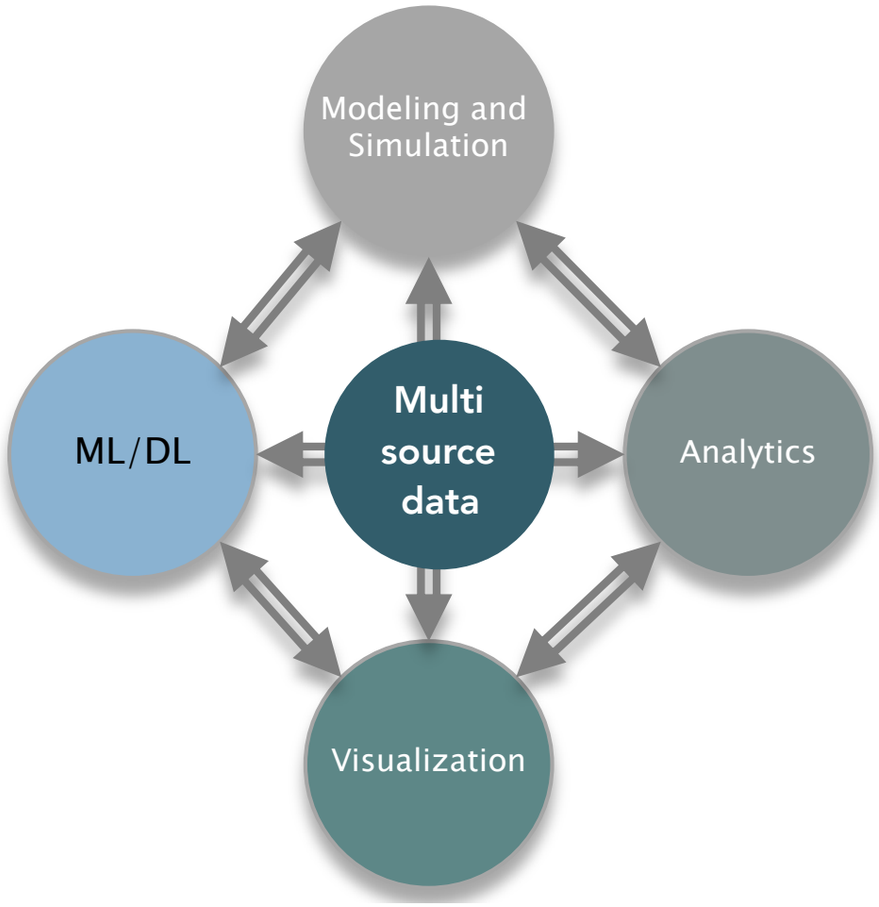
- supported across a continuum of possibly shared infrastructures
- enable science workflows (HDA, HPC, ML) in multi-provider context
- enable data logistics

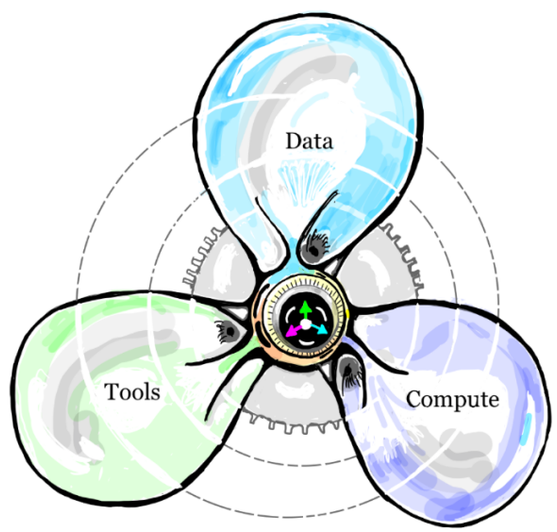
Develop FAIR data policy and management

- aligned with regional & national policies
- including multi-source Open and FAIR data services

Focussed & time-limited working task forces

- e.g. NREN and trans-national data transfers; Data logistics; Software platform of distributed services; Wide-area workflowS; Data life-cycle management and FAIE policy; Users services and support
- co-design approach: SKAO, research users, data and computer scientists, infrastructure developers and providers, funding agencies





Data policy and management

Open data policy

- National observatories (OSU) and observation services (SNO, ANO): data integration/aggregation, standardisation, curation, publication/diffusion (data and derived science products)
- Structured around National infrastructures (IR, TGIR), international organisations (e.g. ESA, ESO), ESFRIs, European and international initiatives (e.g. GoFAIR, EOSC)

Data archiving and curation platforms

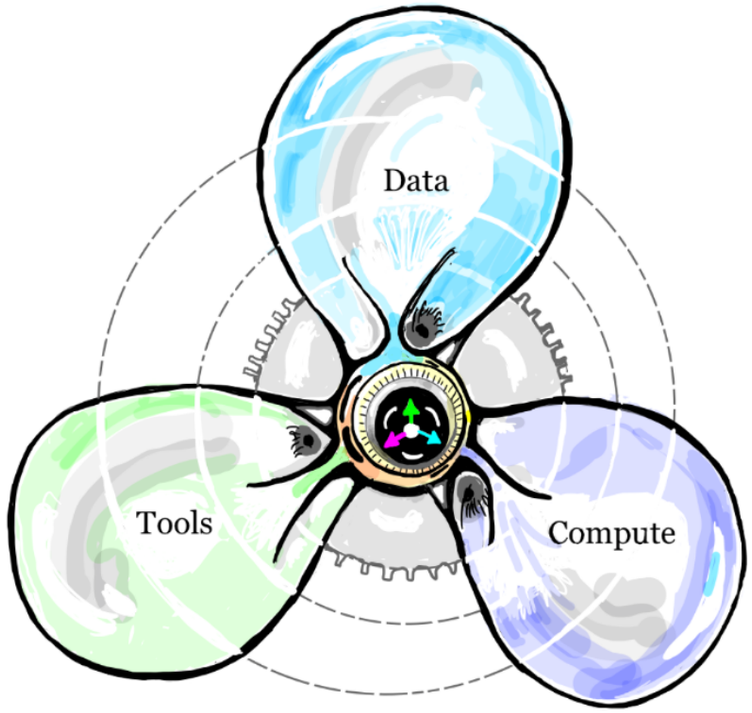
- Linked to long-term programs and instruments
- Mutualise - federate resources, services and critical mass of expertise across disciplines

Thematic poles of multi-source data and services

- National platforms of FAIR services and multi-source data expertise centres: PiDs, registries

Virtual Observatories (e.g. CDS/IVOA, Système Terre, ESGF)

- National and International DOA software platforms of interdisciplinary FAIR data services (translational tools, PiDs, registries, global resolution systems)
- Data model harmonisation (across disciplines) and standardisation (in disciplines)



France SRC : strategic European projects

AENEAS project (PI Michiel van Haarlem, ASTRON)

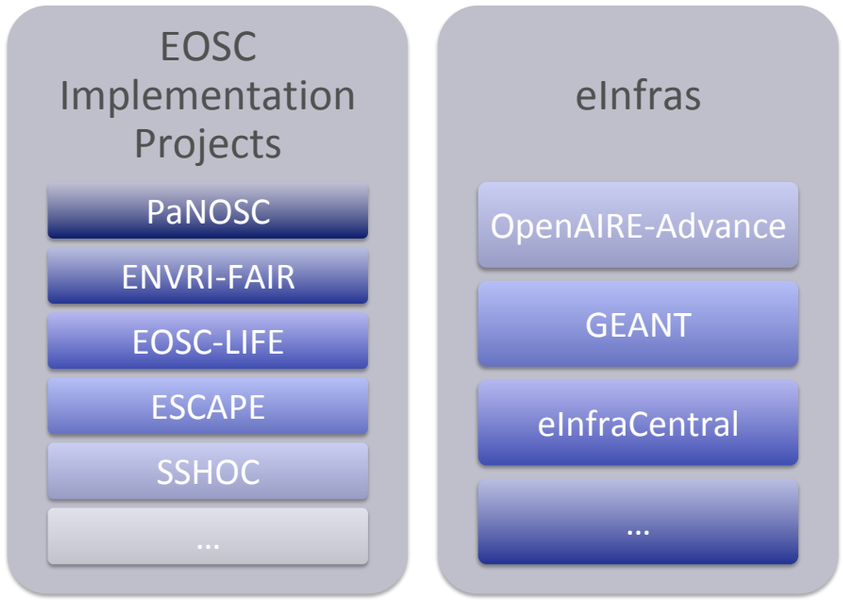
- Design and proof of concept of a distributed European SKA Regional Centre model
- Significant science-driven technical and policy deliverables toward an organisational ESRC model
- Important contribution to an international SRC software platform of distributed services and to an organisational SRC body federating/leveraging a continuum of shared edge and centralised (HPC, Cloud) facilities
- C. Ferrari (OCA, CNRS): general assembly chair ; J.-P. Vilotte (CNRS): member of the EAB (with I. Bird @ CERN and M. Zwaan @ ESO

Aeneas
Advanced European Network of E-infrastructures for Astronomy with the SKA

European Commission | Horizon 2020 European Union funding for Research & Innovation

ESCAPE project (PI Giovanni Lamanna, CNRS)

- European Science cluster of Astronomy & Particle Physics ESFRIs (e.g. CTA, ELT, EST, FAIR, HL-LHC, KM3NeT, SKA) and pan-European RIs (e.g. CERN, ESO, JIV-ERIC, EGO-Virgo)
- Make data and software in multi-messenger astronomy and accelerator-driven particle physics Open and Fair.
- Data and computing services integrated in EOSC Hub, participation to EOSC Hub strategic board
- Contribution to the software platform of distributed services of the ESRC ; coordinated user support and outreach
- Foster collaboration between SKA, ESRC, and IVOA



Relevant European & national context

Europe

- EOSC : EOSC Hub and Pilot projects
- EDI : PRACE-6IP, EUDAT, EGI, CoE H2020, EuroHPC
- ESFRI & European projects : AENEAS, ESCAPE …
- GoFair Initiatives (Netherland, Germany, France) & AGU & RDA Europe
- EXDCI & BDEC initiatives: Big Data and Exascale Computing (demonstrators)

National Level (MESRI)

- National roadmap for research-driven (national and regional) hosting Data and Computing centres (INFRANUM, CODORNUM)
- New national Open Science plan, AI national plan
- ***High-level TGIR committee*** (MESRI): national strategy of large Research Infrastructures
- ***GENCI***: high-performance computing, France contribution (hosting member) to PRACE
- SRC relevant collaborations: CNRS-SKA-France, INRIA, CEA … and industrial partners

CNRS

- MiCaDo: Data and Computing CNRS strategy and implementation; and CNRS-TGIR
- IDRIS and CC-IN2P3: complex data-driven workflows (HPC, HDA, AI)
- Transverse Institut collaborations: IN2P3, INSU …

CNRS-INSU

- Data and High Performance Computing transversal (A&A, OA, ST, IC)
- Software data distributed services platform (multi-messager, IR): CDS/IVOA, Earth Systems …
- ***National data services***: OSUs, SNO, ANO
- ***SKA France***: PPP with private & Industrial partners

Some Challenges ahead

Need formulation of a new vision for

- Service-oriented and Digital Object Architecture
- Software platform of distributed services ***supported*** across a continuum of edge, centralised (HPC, Cloud) and SDCs infrastructures
- **From project-driven to multi messager and multi wavelength science (data reuse)**

AI/ML in data driven Science

- Interpretability, adaptability, physical consistency
- complex noisy data & multi source uncertainty
- Intrinsically limited labels, high-computation
- Interdisciplinary approach & mutualisation

HPC and HDA convergence

- Access policy (FAAI) & security
- Data movement and logistics (from, in, out)
- Resources management and execution environment
- Persistent data storage over data lifecycle
- Data model convergence and FAIR services
- Digital Object Architectures (PiDs, meta data, registries, resolution system)
- FAIR software and librairies lifecycle