From OPEN DATA to OPEN RESEARCH
(new paradigms, challenges and opportunities)

Giovanni Lamanna

LAPP, Laboratoire d'Annecy de Physique des Particules, CNRS-IN2P3

JENAS Symposium – 5 May 2022

@ JENAS 1st Symposium (October 2019) ESCAPE was presented and discussions brought to an important evolution of our work plan: Open Test Science Projects [...]

https://indico.ijclab.in2p3.fr/event/5418/contributions/17542/

- ESCAPE builds an Open Research Environment: current status

https://projectescape.eu

The European Open Science Cloud (EOSC) and the current evolution of the EU Research ecosystem acknowledges the contributions from the Science Clusters and their mutual cooperation

https://indico.in2p3.fr/event/24327/

- Science Cluster scheme as "coordinating structure" for Open Science and operationalising the European Research Area (ERA).

Committing JENA++ scientific communities at large in the Horizon Europe work programme for enabling an operational open science, large cooperative schemes and for Research Infrastructure consolidation

- The new ESCAPE Open Collaboration and its work plan

The ESCAPE H2020 CONSORTIUM

- 31 partners including 2 SMEs

- 10 ESFRI projects & landmarks: CTA, EST, FAIR, HL-LHC, KM3NeT, SKA, LSST, VIRGO, ESO, JIVE

- 2 pan-European International Organizations: CERN, ESO with their world-class established infrastructures, experiments and observatories

- 2 European Research Infrastructures: EGO and JIV-ERIC

- 1 involved initiative/infrastructure: EURO-VO

- 4 supporting European consortia: APPEC, ASTRONET, ECFA and NuPECC

- Budget: **15.98 M€**

- Started: **1/2/2019**

- Duration: **48 months** (end date 31/1/2023)

- Coordinator: **CNRS-LAPP**

Particle Physics / Nuclear Physics

Astrophysics / Astroparticle Physics

Giovanni Lamanna

4

The five Science Clusters

ESCAPE is one of the five Science-Cluster projects that resulted from the H2020 topic call INFRAEOSC-04-2018: *"**Connecting ESFRI infrastructures through Cluster projects**".*
Other Science Clusters: **ENVRI-FAIR** (Environment and Earth Sciences), **EOSC-LIFE** (Biomedical Science), **PANOSC** (Neutron and light sources facilities) and **SSHOC** (Social Science and Humanities).

Expected impact:

- *Improve access to data and tools leading to <u>new insights and innovation</u>*

- *Facilitate access of researchers to data and resources <u>for data driven science</u>.*

- *Create <u>a cross-border</u> open innovation <u>environment</u>.*

- *Rise the efficiency and <u>productivity of researchers</u> through open data services and infrastructures for discovering, accessing, and reusing data.*

- *Foster the establishment of <u>global standards</u>.*

- *Develop <u>synergies</u> and complementarity <u>between involved research infrastructures</u>.*

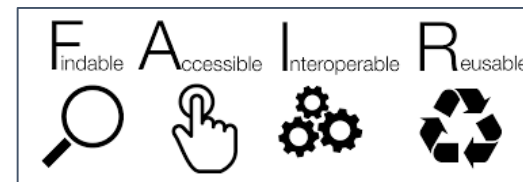- *Adopt <u>common</u> approaches to the <u>data management</u> for economies of scale.*

Working together making data and scientific research FAIR …

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

Five Science Clusters

More than 80% of ESFRI RIs, plus other world-class RIs and new emerging ones.

ESCAPE: Astronomy, Nuclear and Particle Physics ESFRIs

Background analysis for the uptake of "Open Science" and "Data FAIRness"

❑ Builds on communities' complementary excellences in data stewardship:
  - Astronomy Virtual Observatory infrastructure
  - HE-NP expertise in Exabyte-scale data management and large-scale distributed computing

❑ Builds on existing inter-RI synergies, intersections.

❑ Recognises that ESCAPE communities will be Exascale data generators, early adopters of ICT and data management innovations, push state-of-the-art.

❑ Both Observatory- and Facility- operations require global, open access to data, long term curation, and sustainability.

❑ Concerned by connecting the ESFRIs to EOSC - European Open Science Cloud (at the early stage when EOSC was just a concept and EU shared policy)

ESCAPE consortium

As per H2020 INFRAEOSC-04-2018 call - CLUSTER MEMBERSHIP and PARTNERSHIP:

- *The EC funding contributions **proportional to the number** of pan-European research infrastructures (**ESFRI project/landmark)** that the science cluster connects to the EOSC.*

- Each **RI legal entity commits** together with a sub-set of <u>associated national institutes</u>.

Furthermore*:*

- The <u>Director</u> of each ESFRI RI is a member of the **ESCAPE Supervisory Committee (E-SC)**

- <u>APPEC, ASTRONET, ECFA, NuPPEC</u> chairs and <u>ESA</u> representative form the **ESCAPE External Advisory Board (E-EAB)**
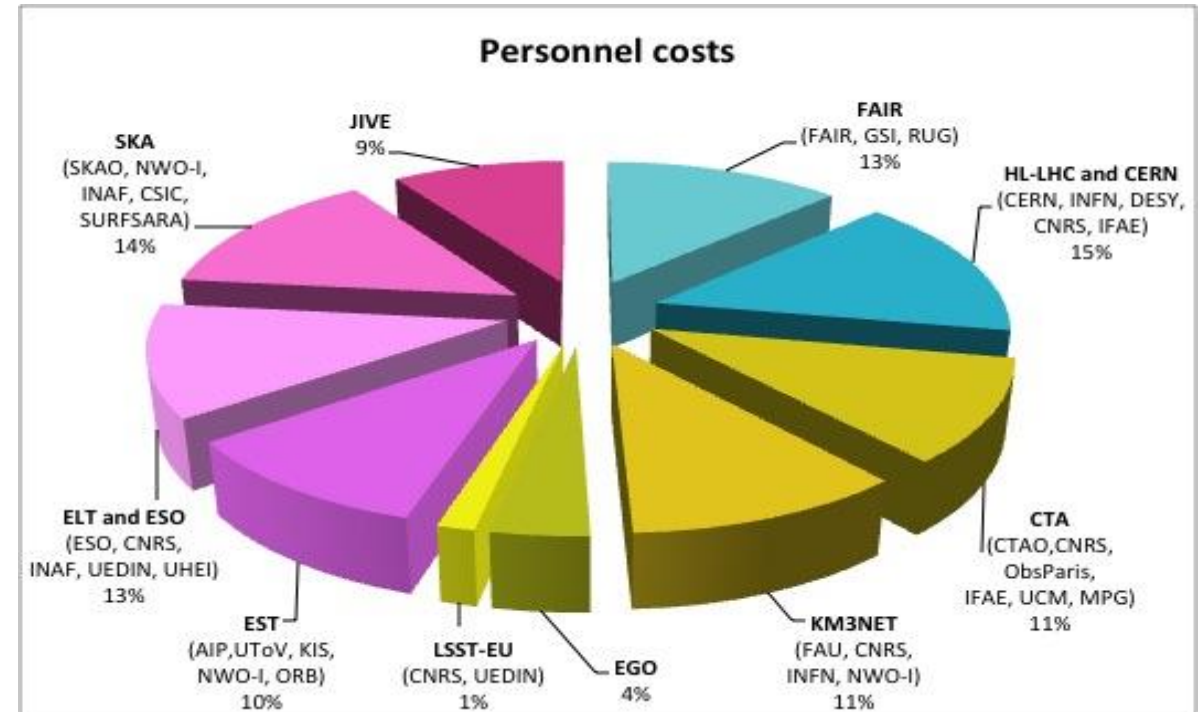
Personnel costs

Distribution of personnel costs among partners grouped by RI

WORK PROGRAMME

Data Lake:
Build a scalable, federated, data infrastructure as the basis of open science for the ESFRI projects within ESCAPE.

Software Repository:
Repository of "scientific software" as a major component of the "data" to be curated in EOSC.
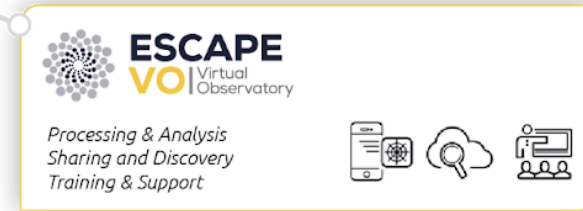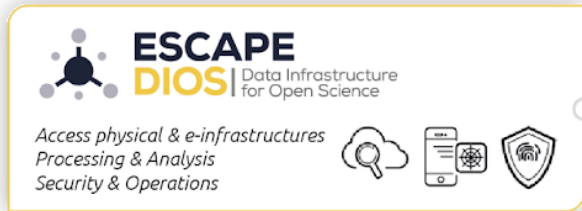
Science Platforms:
Flexible science platforms to enable the open data analysis tailored by and for each facility as well as a global one for transversal workflows.

Citizen Science:
Open gateway for citizen science on ESCAPE data archives and ESFRI community

Virtual Observatory:
Extend the VO *FAIR* standards, methods and to a broader scientific context; prepare the VO to interface the large data volumes of next facilities.

Our domain-based "EOSC CELL"

Overview progress: implementing the EOSC-cell

From preliminary plans or first prototypes

to operational components of the ESCAPE EOSC-cell for open science:

The ESCAPE Data Infrastructure for Open Science (DIOS)

- The main **Data Lake (DL)** building blocks are deployed and **operated by the ESCAPE partner institutes**.

- **Petabyte scale storage** harnessing capacity provided at DESY, SURF-SARA, IN2P3-CC, CERN, IFAE-PIC, LAPP, GSI and INFN (CNAF, ROMA and Napoli)

- **Data management and storage orchestration** layer provided by Rucio (framework developed in ATLAS/LHC)

- **File transfer and data movement services** provided by FTS (developed in WLCG/CERN) and **enabling different transfer and access protocols**: http, gridftp, xrootd and swift/S3

- **Global Data Lake Information System**: endpoints, protocols, ports provided by CRIC (from ATLAS/LHC)

- **Common Auth/Authz/IM (AAI) - the ESCAPE IAM**, from INDIGO-DATACLOUD, AARC/AARC2 and FIM4R projects, deployed and operated in INFN/CNAF
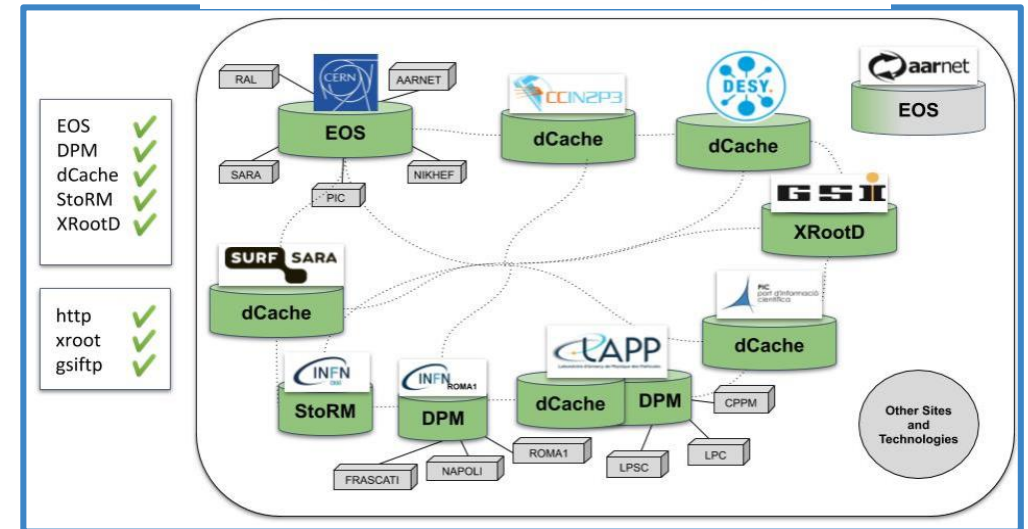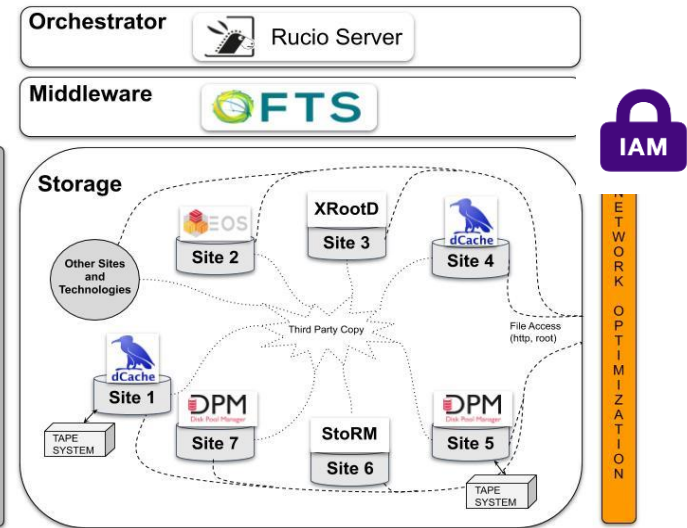
Overview progress: implementing the EOSC-cell

## *A functional Data Lake implemented* (by all RIs and for their requirements)

The ESCAPE pilot **Data Lake** (DL), after design and deployment was recently assessed, culminating in a joint exercise labelled as **Full Dress Rehearsal** (FDR20):
- a 24h production-like window where ESFRI's executed relevant workloads.

Followed by a full scale **Data and Analysis Challenge** (DAC21) performing production-like Data Management, Processing and Analysis workloads including interplay possibilities using large scale resources (batch systems and clouds) and user-analysis oriented platforms (online notebooks and analysis platforms).
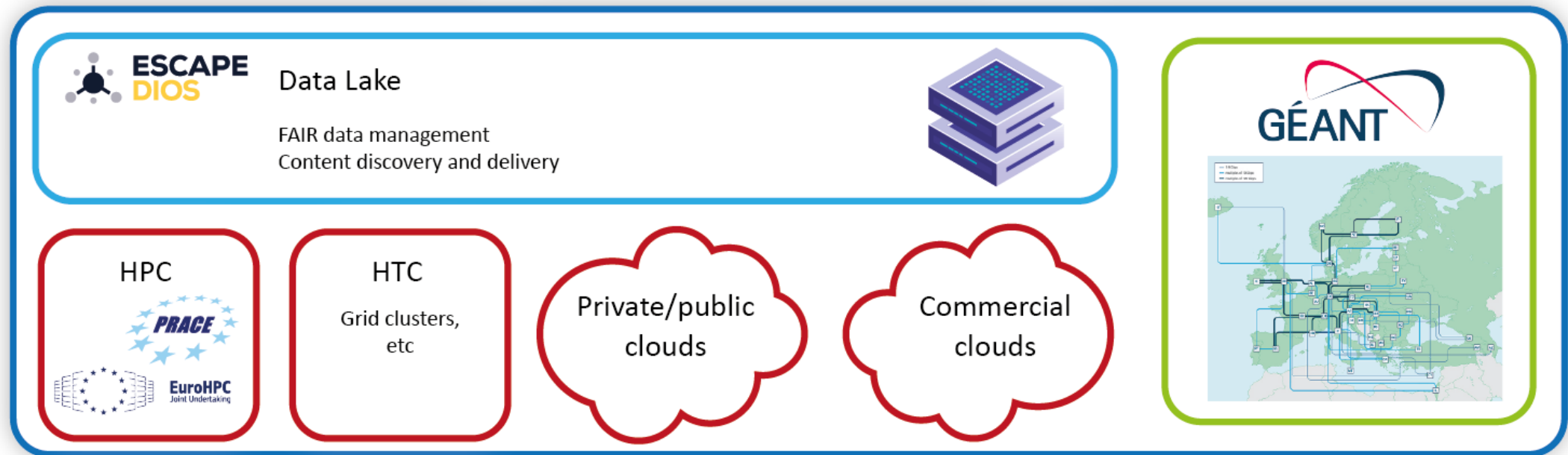
# The ESCAPE Data Infrastructure for Open Science (DIOS)
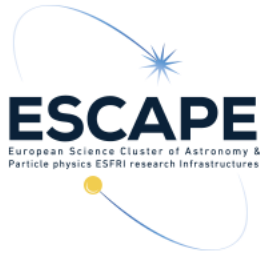
*Raw data injected, stored and preserved in the DL. Data processed by users, results are stored back in the DL.*

Offload data from the storage buffer in the coast, replicate across sites, run data calibration, store back. Data product ready for user consumption

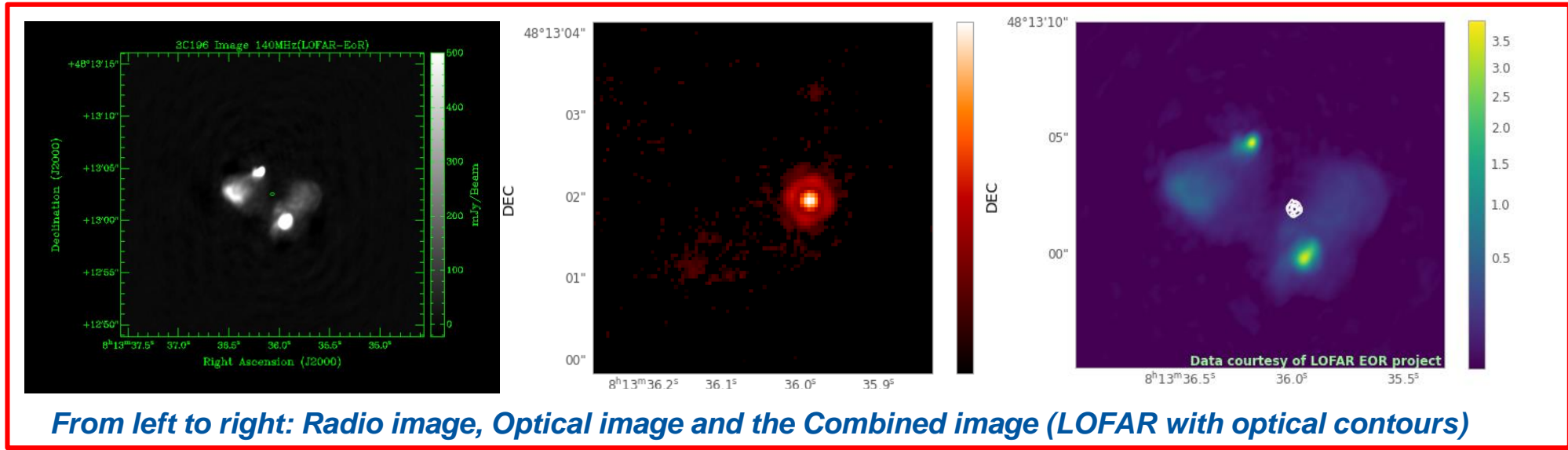*Distributed data re-processing taken at remote locations (La Palma)*

Large experiment demonstrating open data capabilities (with ad-hoc ATLAS open-data software)

The ESCAPE Data Infrastructure for Open Science (DIOS)
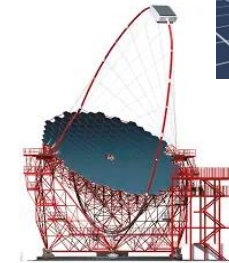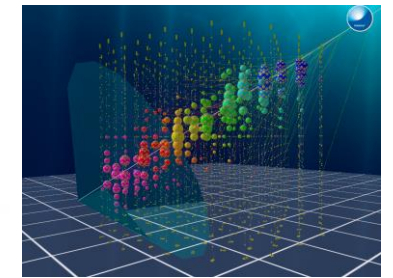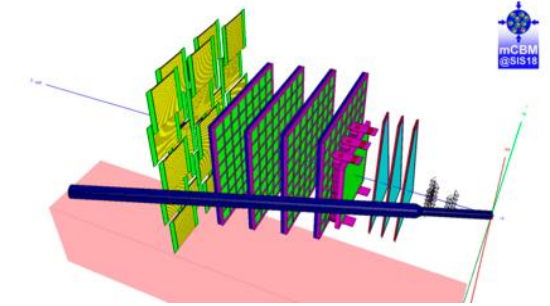
Full-cycle scientific data management and data processing

| 1. Data injected to the DL from **three** radio source observations in external locations | 2. User in external location download the data, process and store results back to the DL | 3. User interested in combining results stored with other public data to cover also visible spectrum | 4. Combined optical data from the Hubble located via the **VO (WP4)** | 5. Optical and radio data aggregate in via the **ESAP (WP5), combined analysis done.** Results uploaded back to the DL. |

From left to right: Radio image, Optical image and the Combined image (LOFAR with optical contours)

The ESCAPE Data Infrastructure for Open Science (DIOS)

Full-cycle scientific data management and data processing

Global-scale (Australia, South-Africa and Northern hemisphere sites)Data Management

Simulate observations and leverage telescope local storage data replication to fulfil daily data management cycles (incorporating SLAC in the data replication chain)

DL interface with local and heterogeneous resources (HPC), control access, caching and analysis

Overview progress: implementing the EOSC-cell

Deployment of the Open-source ESCAPE catalogue of resources (data and software)

Shared domain-based open science software and services based on FAIR principles

Objectives:

- Facilitate and support continuous **development, deployment, exposure and preservation** of partners' software/tools/services

- Foster **interoperability, software re-use and cross-fertilisation** between ESFRIs (e.g. simulation)

- Offer an **open innovation environment for open standards** (e.g. workflows, data-formats), **common regulations** and **shared (novel) software** for multi-messenger & multi-probe data

eOSSR library https://gitlab.in2p3.fr/escape2020/wp3/eossr

- incorporates all OSSR developments, based on the commonly defined practices and standards
- python-based
- OSSR API : send request to the OSSR, find and filter software and services, upload new entries, update existing entries
- CI : automated upload / update using gitlab CI
- Metadata : schema definition, crosswalk between CodeMeta and Zenodo, generator and validator available

ESCAPE OSSR

Catalogue & Repository of resources

Datasets
Software & services
Tutorials
Training
Publications

eOSSR

DOI 10.5281/zenodo.5524912

publishes source code (updates your existing record with new versions)

- long term archived
- findable
- citable

GitLab

- develop/maintain software
- tag a version(release)
- add metadata
- let the CI do the rest

- builds a container image

- publishes image

- registers image

- integrates into ESCAPE EOSC cell

❑Software as first class EOSC citizen

❑Enable software custodians

❑All lectures/materials online:

https://escape2020.github.io/school2021/

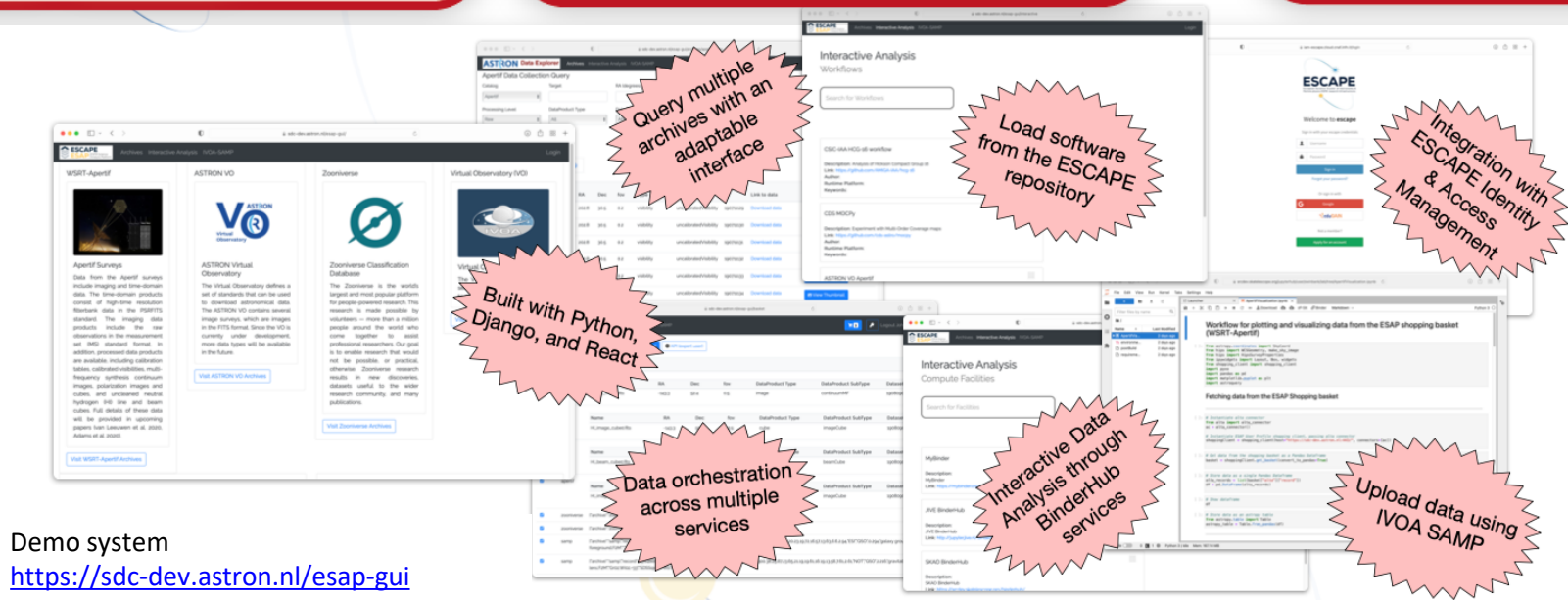
https://indico.in2p3.fr/event/26913/

Overview progress: implementing the EOSC-cell

- Development of the **ESFRI Science Analysis Platform**: a toolkit for building platforms through which users can discover and interact with the data products, software tools, workflows, and services that are made available through ESCAPE.
- Preparing ESFRI services, data products, and tools for integration with ESAP and their subsequent use within ESCAPE and EOSC.

Overview progress: implementing the EOSC-cell

Demo system
https://sdc-dev.astron.nl/esap-gui

Overview progress: implementing the EOSC-cell

Full integration of ESAP with the ESCAPE Zooniverse "Panoptes" system.

Giovanni Lamanna

Overview progress: implementing the EOSC-cell

TSP's

RI-Specific Science Platforms

ESCAPE VO Virtual Observatory

Astronomy Data centres	VO Registry

VO Registry
Analysis Tools
VO Services

ESCAPE SAP Science Platforms

Workflows, notebooks, deployment platforms, packaging

ESCAPE CS Citizen Science

Enables a *Virtual Research Environment* of interoperable tools and services based on IVOA standards, including ESO Science Archive and ESO Science Portal
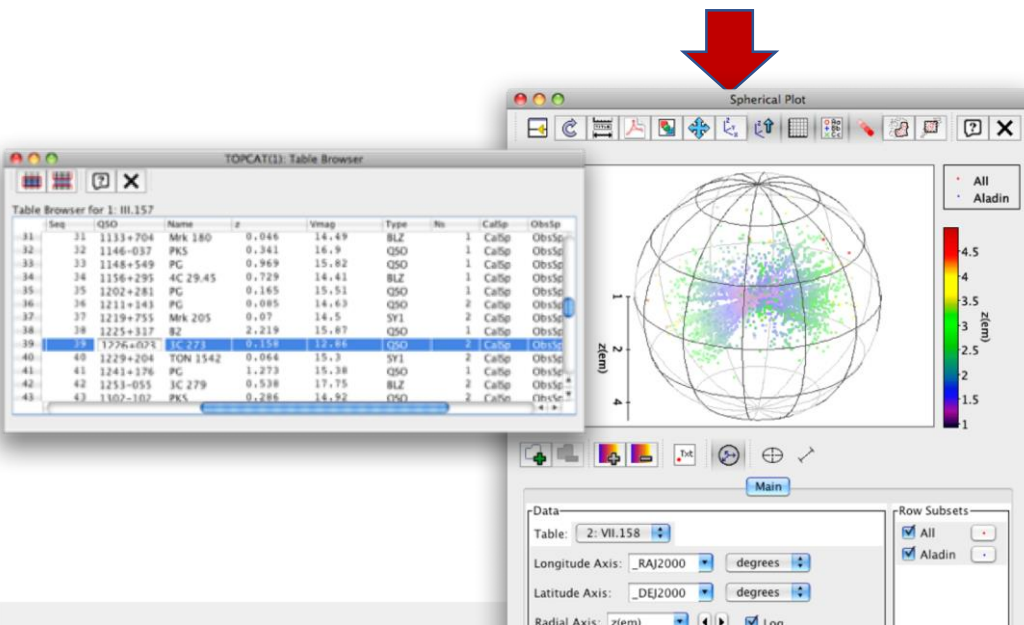
Implementing the ESCAPE disciplinary EOSC-cell

The ESCAPE EOSC-cell for a three-fold impact:

- a FAIR digital environment dedicated to scientists at large to interoperate data and workflows for fundamental science;
- data management solutions mutually adopted by a large fraction of RIs and potentially extendible to further disciplines;
- the "**EOSC Web of FAIR Data and Services for Science**" for our disciplines

➡ A **Virtual Research Environment (VRE):** thematic collaborative digital environment used by scientists, which enables FAIR community-based scientific research, training, innovation, cross-fertilisation and open science.

Evolving practices on the assessment of research giving increasing value to open science contributions and outputs beyond publications. A wide range of digital objects beyond publications, including data, software, code, workflows, and processes, such as open peer-reviews (requiring an enhanced traceability, coherent and comprehensive metrics and FAIRness of a wide range of digital objects). Digital content added in order to :

- Perform analysis
- Explore analysis
- Repeat analysis
- Modify analysis
- Upload analysis
- Publish new results
- Rewarding scientists

**The ESCAPE Virtual Research Environment prototype
to host the Test Science Projects**
(and part of the EOSC Future project [...])

(Test) Science Projects – (T)SPs

Originally part of the ESCAPE work programme, proposed to validate ESCAPE services for Open Science at the end of the project.

The ESCAPE-TSP concept finds consensus and evolves for a larger impact.

- Stimulating and/or cooperating with Joint ECFA, NuPPEC, APPEC Activities (JENAA)

- TSP "bench" concept aiming at enhancing researchers participation in open science and cross-domain scientific research (and guiding the EOSC architecture).

- Included in the new H2020 EOSC Future project

Synergy between Clusters & e- infrastructures
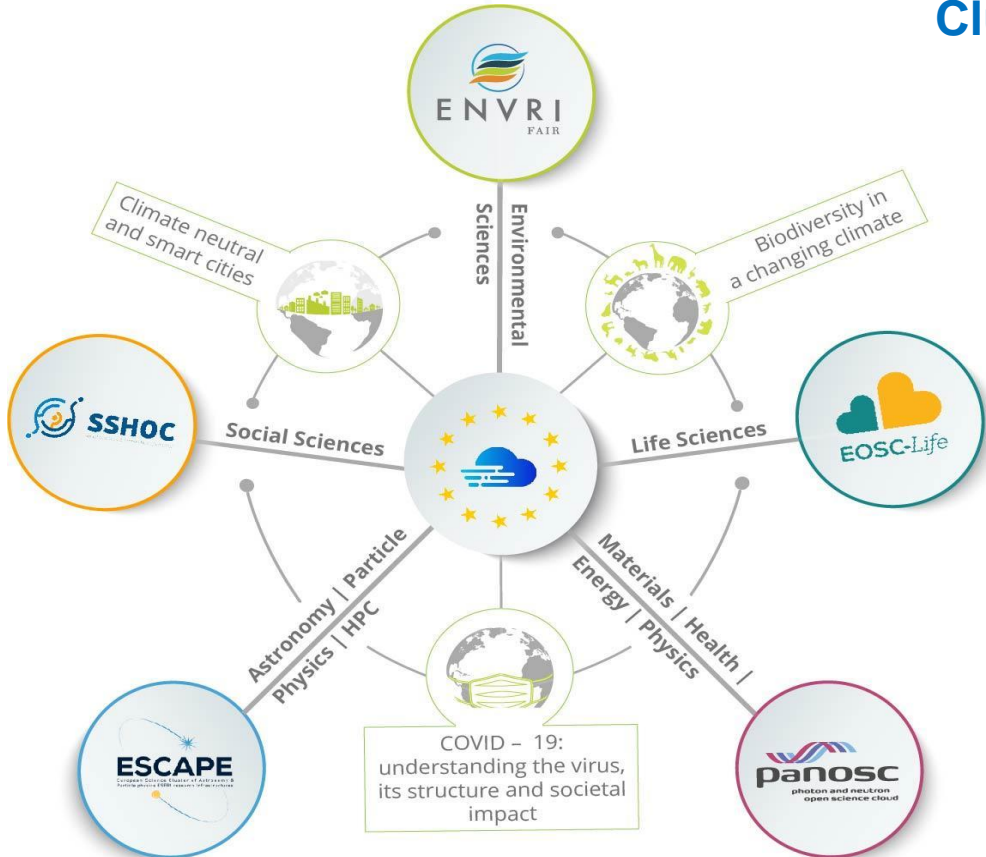
- EOSC Service Delivery
- Innovation capacity and procurement
- Architecture and Interoperability
- Design and Development of Portal Layers
- Training and Skills
- Integration of Community Services and Products
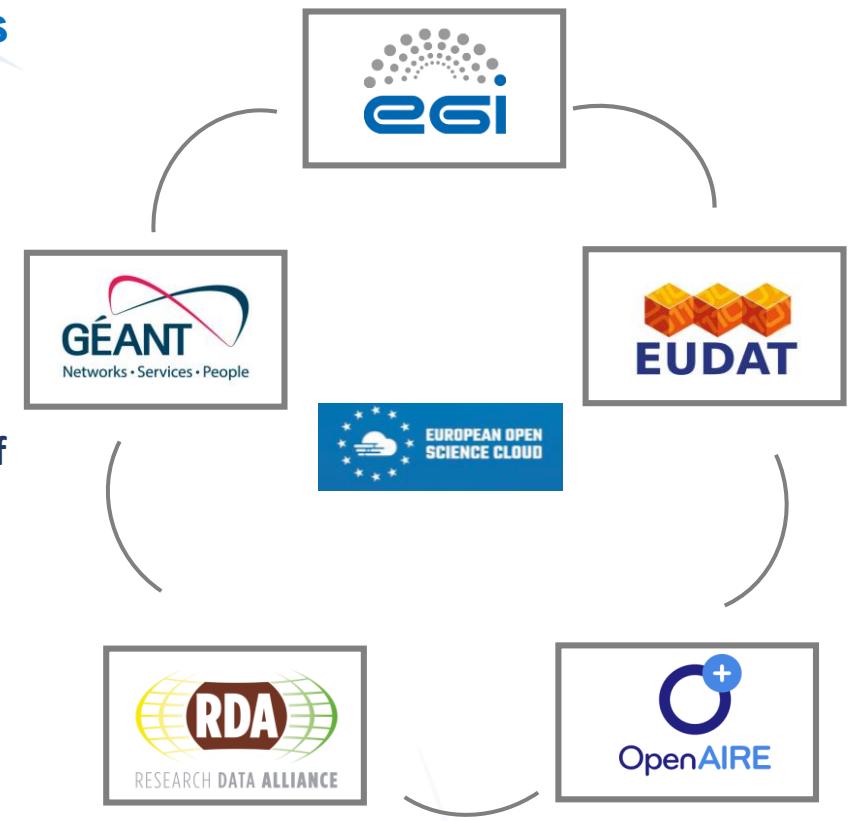
TSPs are proposed to demonstrate multi-domain science integration across ESCAPE

- demonstrate new cutting edge open science capabilities, making use of the services implemented within ESCAPE

- feedback on the capabilities delivered by ESCAPE

- benefit real science goals in exploring synergies between the ESFRIs and largely among three scientific communities Astrophysics/Astroparticle, accelerator-based Particle and Nuclear Physics (supported by consortia of EU member states research agencies and institutes within JENAA)

1. Dark Matter Science Project
2. Extreme Universe (& Gravitational waves) Science Project

Linked to two corresponding JENAA EoIs
(with already about 1000 subscribed scientists)
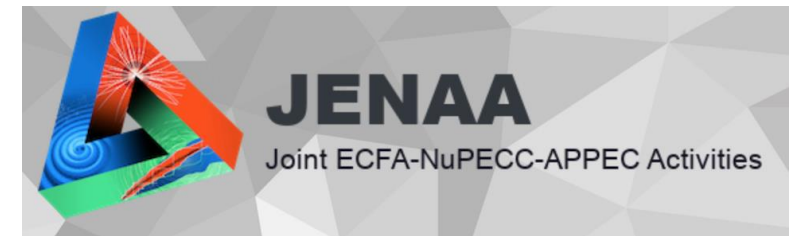
A top-down endorsement for a bottom-up approach based on Expression of Interests (EoI) subscribed by researchers

Open Science Pilots within "Dark Matter" as focus / scientific question

Primary goal of TSP: produce **summary plots** for different hypotheses (starting from WIMP hypothesis) using fully reproducible workflows on the **EOSC**, using **ESCAPE services**
- Starting for now from RIs in ESCAPE partners (LHC, CTA, KM3Net, Darkside …) , will open up to others once the first workflows are up
- **Bottom-up** effort: start from the individual science inputs and make tools + digital objects (data) shareable and sustainable

Additional goal of TSP: share software methods that can be useful to other communities (e.g. ML algorithms)
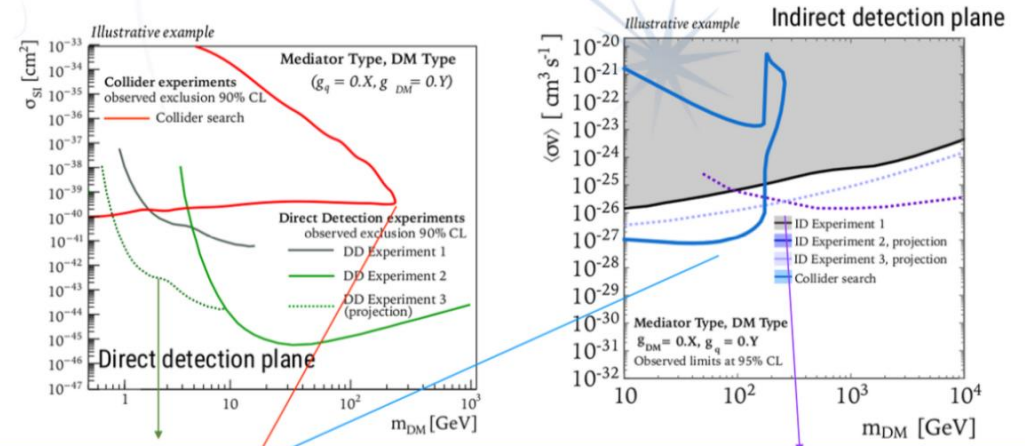
METHODS	Direct Detection	Colliders	Astrophysics	Theory	Indirect Detection
RELEVANCE	DM that interacts inside the detector (WIMPs, axions)	produce DM and probe the dark interaction	necessary for all	necessary for all	detect annihilating/ decaying DM through its decays (i.e. neutrino searches, gamma rays)
EXPERIMENTS INVOLVED	Darkside	ATLAS			KM3Net, CTA

Dark Matter Science Project

Ex. LHC. Work in progress on:

- **Individual DM analyses (science cases)**
  - For examples from colliders: see Jared Little and Josh Greaves's JENAS posters

- Preparing and onboarding **software**, discussing **data access**

- **Virtual Research Environment (VRE)**
  - Building the core of the project where the workflows will run

- Understanding **computing resource needs**

First full data analysis results expected for October 2022

J. Greaves (with P. Harris and C. Doglioni): reinterpreting ATLAS DM results to connect to accelerator DM searches
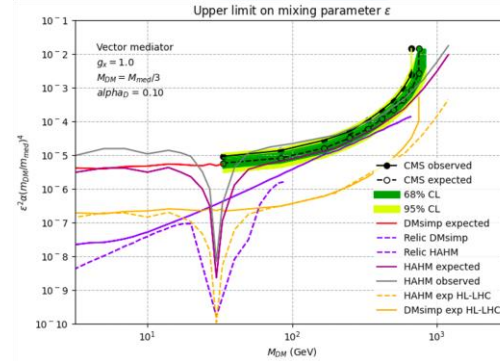
J. Little (with T. Hryn'ova): making DM summary plots using ATLAS High Luminosity projections, withinthe ESCAPE VRE

E. Gazzarrini, E. Garcia

Extreme Universe Science Project
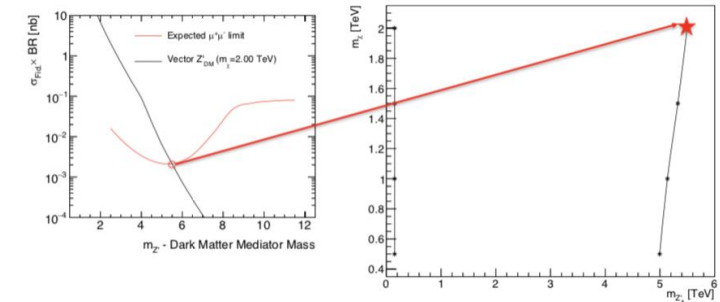
A series of pilots focused on violent phenomena in the Universe with **Astrophyiscal** as well as **fundamental implications** (e.g. Dark Matter)

Collecting requirements for VRE.

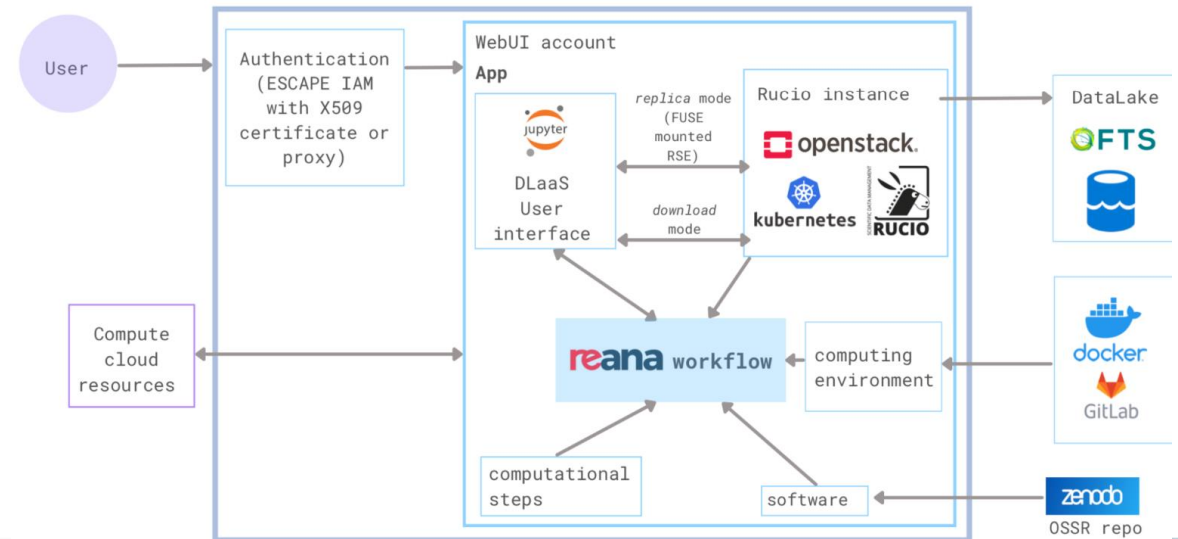**Understanding services, computing resource needs and technical challenges.**

First full data analysis results expected for October 2022

Main Research Area	Objects/sources	Messengers	ESF/RI involved	ESCAPE services EOSC-Future integrations	Data Analysis tools (AI,ML)	Pilot project(s)	Computing resources required
Compact objects	*Pulsars, FRBs, Off-nuclear AGN*	*radio, optical, X-ray, ...*	*LOFAR...*	*Multiwavelength platform/Software catalogue,VO tools*	*Data science, Machine Learning*	*1) Radio astronomy: FRBs, pulsars, plerions, off-nuclear AGN*	*Compute cluster, Jupyter hub, Rucio Data lake*
High energy Astrophysics	*GRBs, jets, AGN, BNS, CCSN*	*neutrinos, gamma-ray, radio,X-ray, GW,...*	*CTA, Virgo, KM3NeT, SKA,LSST*	*Multimessenger platform/Software catalogue,... Virtual Observatory tools*	*Model comparison, Machine Learning*	*1)GRB/neutrino/GW analysis, 2) Blazar MWL/neutrino*	*GPU cluster Jupyter hub*
Fundamental physics	*Dark matter, GR, Primordial Universe*	*GW,*	*Virgo, Einstein Telescope*	*Template banks, generation software,...*	*Machine learning approach*	*1) DM template bank and ML analysis pipeline*	*GPU cluster Jupyter hub*

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

Extreme Universe Science Project

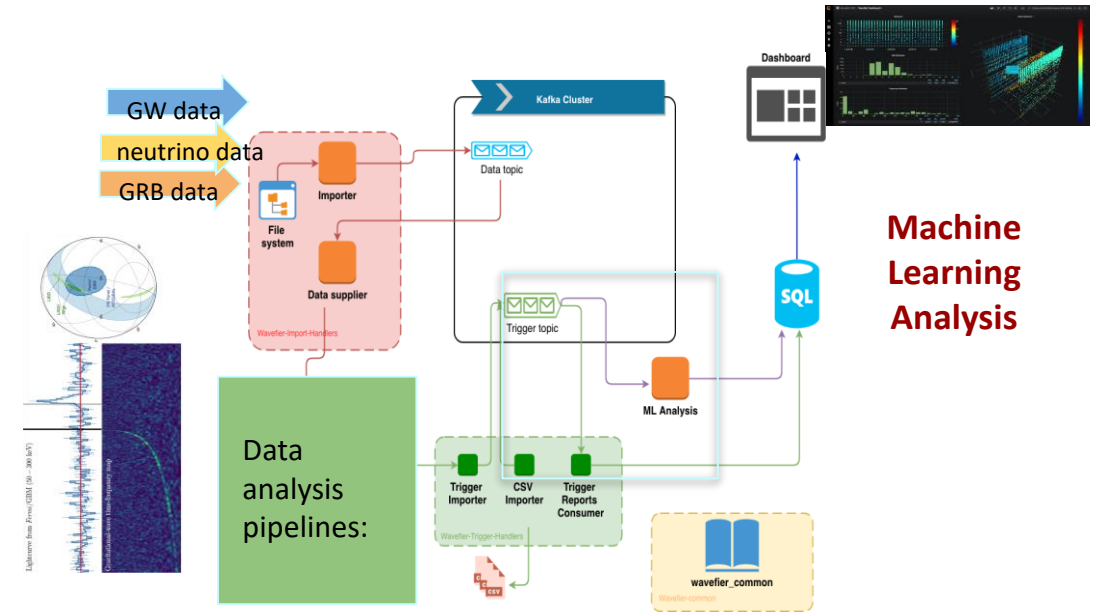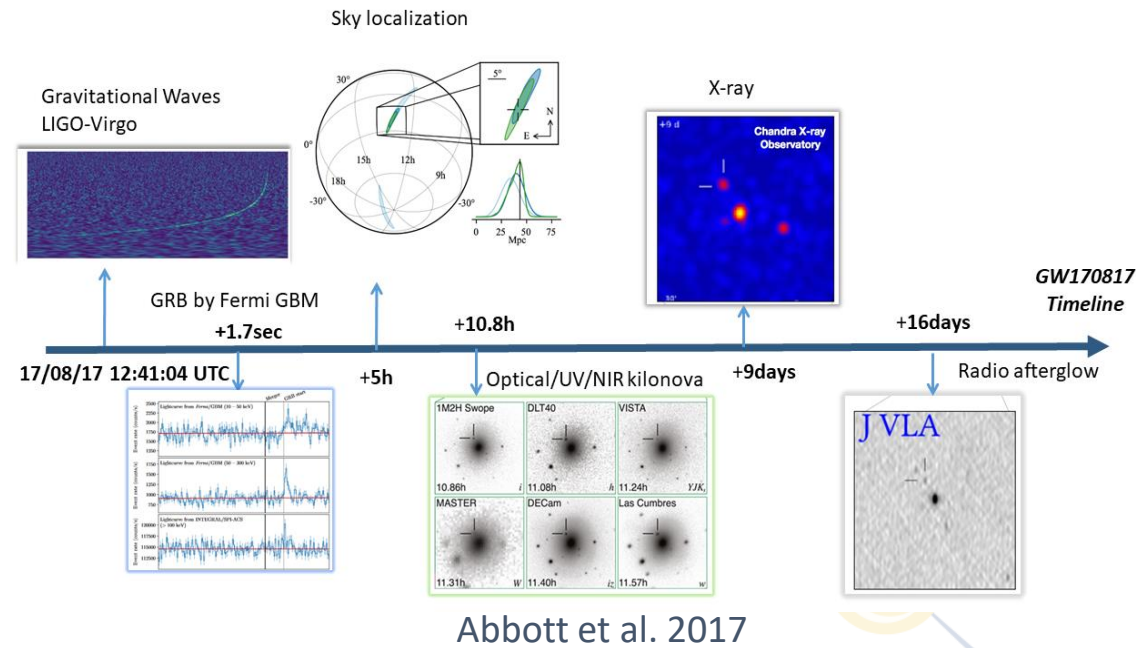## Multi-Messenger Astrophysics

⓾ **Analyzing Multimessenger signals within ESCAPE/EOSC framework with a real time transient signal classifier**

Abbott et al. 2017

Wavefier: A prototype for Real time analysis

- *Multimessenger platform/Software catalogue*
- *GRB/neutrino/GW analysis*
- *EOSC implementation of real time classifier*

**Machine Learning Analysis**

Elena Cuoco

Fundamental physics (.. again Dark Matter) by emerging topics (e.g. Einstein Telescope perpsectives)

Template banks, generation software
DM template bank and ML analysis pipeline

- Frequency band of ET means that most promising target would be solar and sub-solar mass binaries

- **Primordial black holes** (PBHs) could form such binaries, and must be surrounded by a dense spike of particle DM

- Currently developing an ET pipeline to search for 'de-phased' GW waveforms due to DM around PBHs, using machine learning

- Waveform generation & search pipeline will all be public → implementation in virtual research environment will allow easy **access** and **re-use**
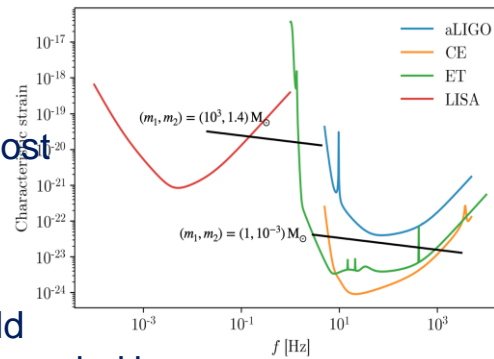
Distinguish DM matter presence with GW detection and waveform classification

Possibility to detect and constrain dense DM 'spikes' with just a few cycles of GW 'dephasing' → but these subtle differences can be hard to detect!

Kavanagh et al., https://arxiv.org/abs/2002.12811
Coogan et al., https://arxiv.org/abs/2002.12811
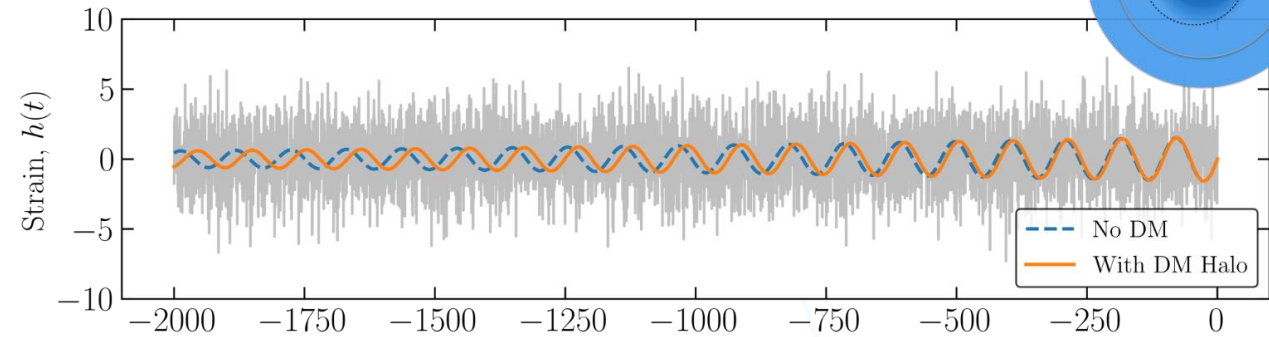
Science Cluster synergies and outlook for the future
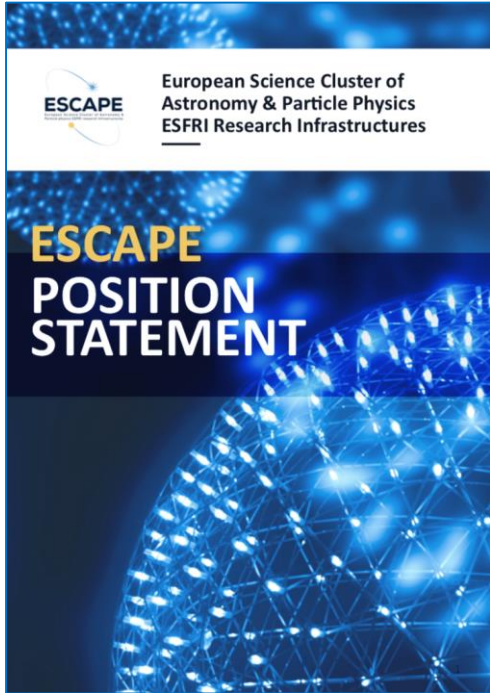
https://www.projectesca
pe.eu/sites/default/files/
Escape_position_stateme
nt_web.pdf

https://zenodo.org/record/3675081 - .X2R2PJNLhTY

https://zenodo.org/record/4889503

Science Cluster synergies and outlook for the future

The Science Clusters occupy a unique position between EOSC, ESFRI RIs and scientific communities.
Three momenta mark the success of the **Science Clusters** -> We all want to keep on them for the future.

Top-Down: The **(ESFRI) RIs** legal entities ➡️ joining efforts together

Bottom-Up: The **concerned scientists** ➡️ willing to pursue the cross-fertilization in science and innovation

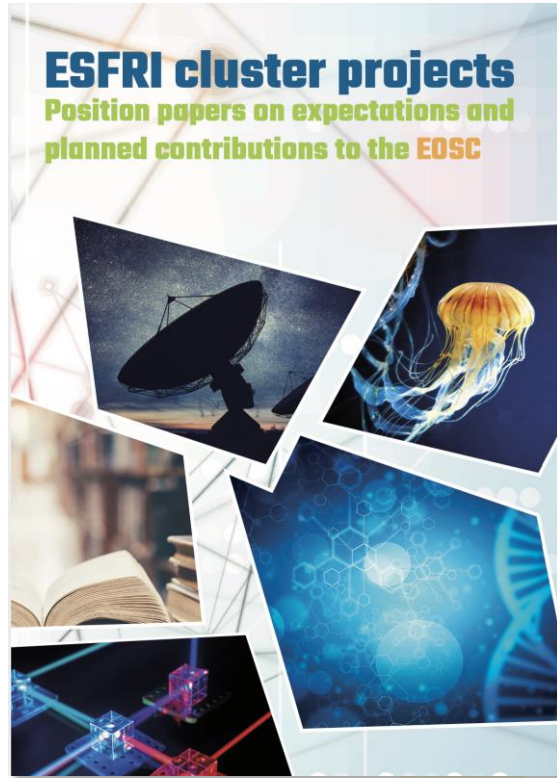**Horizontally:** The Universities and Institutes ➡️ leveraging the inter-domain potential… to be fully exploited around new academic/training schemes based on data-research

The five Science Clusters have debated and positioned their own community-based expectations in the Horizon Europe perspective.

-> they are moving towards sustained platforms/collaborations

https://indico.in2p3.fr/event/24327/

ESCAPE will become a sustained "Community Platform".

Its (ESF)RIs core partners as Parties in a new **Collaboration Agreement**

to operate Open Science as well as cooperating in order to address new topics in ERA.

[...]

Giovanni Lamanna

New ESCAPE Collaboration Agreement

ESCAPE collaboration preliminary Work Plan:

For EOSC destination and European Research Area (ERA) operational

- Data Lake
- Software catalogue, methods and training
- European Virtual Software Institute for Research
- Virtual Research Environments
- Open Science Research Projects
- Data interoperability

- Low Latency data analysis
- Technological dev. and innovation (e.g. quantum)
- Curiosity driven RIs' synergies
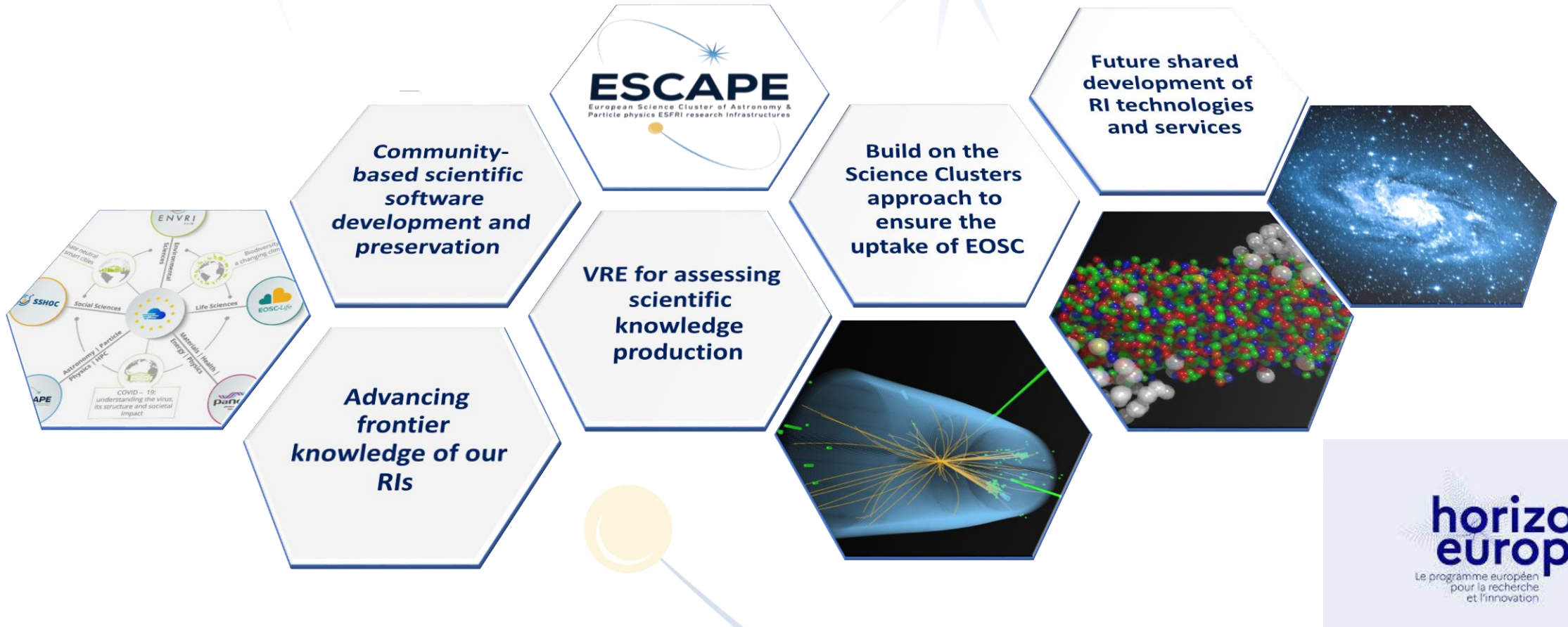- Inter-cluster projects for society
- Citizen Science
- Researchers' rewarding
- Sector Data Spaces […]

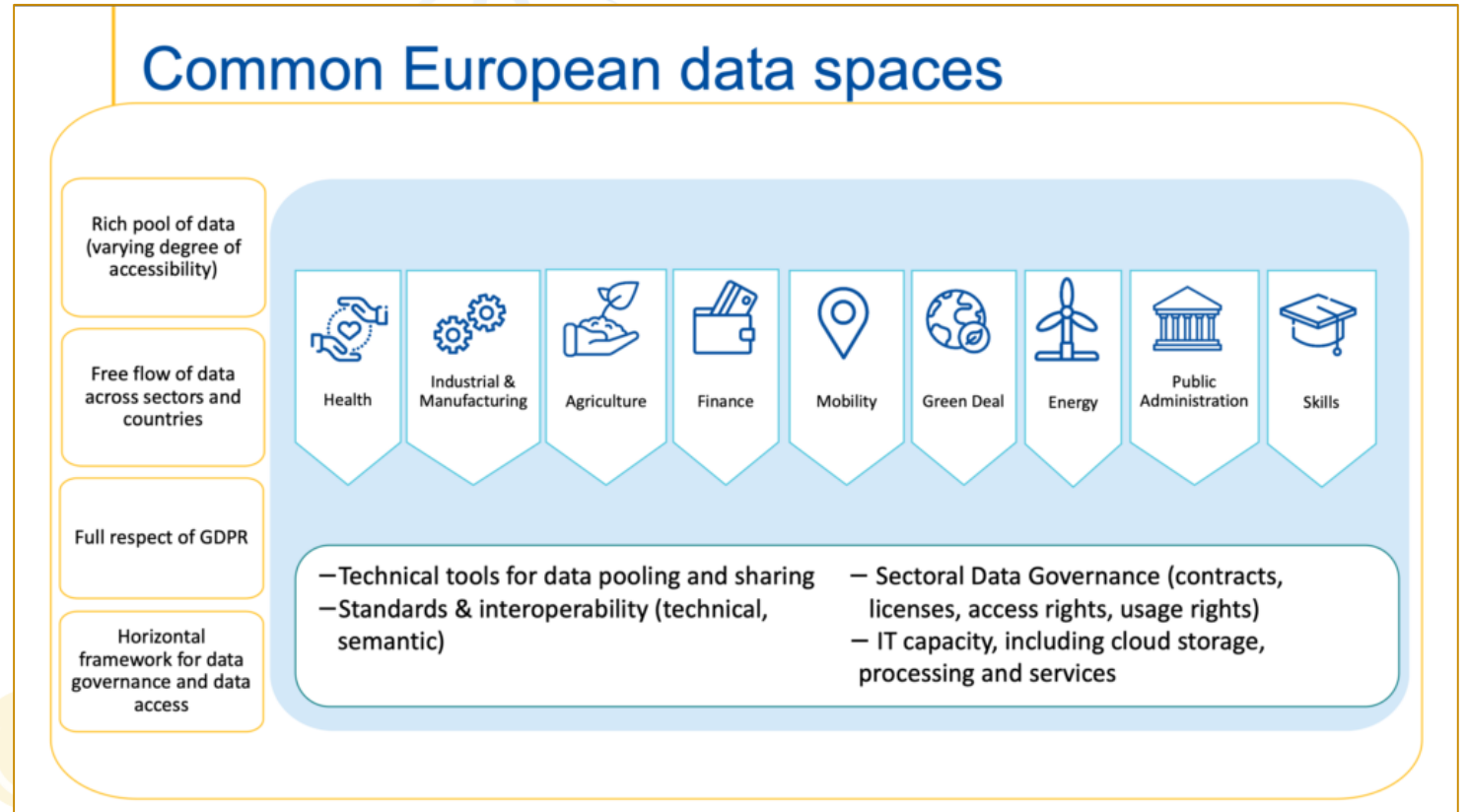# The ESCAPE collaboration work plan in Horizon Europe

Matching the five Science Clusters expectations as well as the new ESCAPE work plan

European Data Spaces: "…The European strategy for data aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty.

Common European data spaces will ensure that more data becomes available for use in the economy and society, while keeping companies and individuals who generate the data in control.

Data is an essential resource for economic growth, competitiveness, innovation, job creation and societal progress in general…"

Common European data spaces

Rich pool of data (varying degree of accessibility)

Free flow of data across sectors and countries

Full respect of GDPR

Horizontal framework for data governance and data access

Health | Industrial & Manufacturing | Agriculture | Finance | Mobility | Green Deal | Energy | Public Administration | Skills

— Technical tools for data pooling and sharing
— Standards & interoperability (technical, semantic)

— Sectoral Data Governance (contracts, licenses, access rights, usage rights)
— IT capacity, including cloud storage, processing and services

Last outlook: sector data space

…The ESCAPE Big-Science RIs (as well as all other Science Clusters) are concerned by the European Data Strategy…

Industrial & Manufacturing

ESFRI facilities to support EU's industry; as well as global coordination for innovation, e.g. Quantum Computing as a service and for tech. R&D; algorithms and standards for Industry 4.0; AI; HPC.

Green Deal

Aligning Innovation programmes and technology R&D with environmental implications.

Energy

Cross-sector sharing of data for an unifying, forward-looking approach of any Big Science facility for energy efficiency, water management, etc.

Public Administration

Finance

Data and experience to improve accountability of public spending for research; management and sustainability of Big Science RIs through socio-economic impact model analyses.

European Strategy for Data
A common European data space, a single market for data

- Data can flow within the EU and across sectors
- Availability of high quality data to create and innovate
- European rules and values are fully respected
- Rules for access and use of data are fair, practical and clear & clear data governance mechanisms are in place

European Commission

Skills

Science Data to reduce the skills mismatches between the education and training systems and the labour market needs.

Health

Health data essential for EOSC-Life advances but also link with particle/nuclear physics facilities applied to healthcare

Conclusions

"OPEN DATA and OPEN RESEARCH", a change of the global paradigm.

ESCAPE and other Science Clusters position:

Expectation:

Accelerate the discoveries and increase scientific value

Enable opportunities offered by the digital revolution

Commitment:

Shape and operate platforms for data interoperability

Sustain the federation of RIs for excellence science

Widen impacts of Open (Data) Science

Cooperation:

EOSC Association and EC

Inter-domain and cross-disciplines

SMEs for co-developments

Society at large

Inclusiveness:

Attracting more thematic and emerging RIs

Lead a regional as well as international alliance in Science

Universities, Institutes and citizens

ESCAPE
European Science Cluster of Astronomy & Particle physics ESFRI research Infrastructures

THANK YOU!