



ESCAPE

European Science Cluster of Astronomy &
Particle physics ESFRI research Infrastructures

ESCAPE DIOS: the final meeting

Xavier Espinal (CERN)

Final Meeting, Amsterdam, 7-8th June 2022





Ceci n'est pas une pipe.

This is **NOT** A_(conventional) **MEETING**

- **Presentation-free** meeting! ⁽¹⁾
- Forums-workgroups
- **Plenty of time** for coffee/breakouts **discussions**
- Start/end times of the forums **indicative**
- It is **fundamental to document** the outcomes of our discussions to consolidate ideas and solidify initiatives, this is a duty for everyone, please contribute to the:
 - **Live document**

Ceci n'est pas une pipe.

This is **NOT** A_(conventional) **MEETING**

- **Presentation-free** meeting! ⁽¹⁾ **To exchange** rather than only listen
- Forums-workgroups
- **Plenty of time** for coffee/breakouts **discussions**
- Start/end times of the forums are **indicative**
 - Feel free to bump in (and out)
- It is **fundamental to document** the outcomes of our discussions to consolidate ideas and solidify initiatives, this is a duty for all of us, please contribute to the:

Live document

⁽¹⁾ Except me

Ceci n'est pas une pipe.

Meeting Agenda: Tuesday

TUESDAY, JUNE 7



2:00 PM → 2:30 PM **Welcome** 🕒 30m

2:30 PM → 3:15 PM **The ESCAPE DIOS experience** 🕒 45m

Speakers: Xavier Espinal (CERN), Rosie Bolton (Square Kilometre Array Organisation)

3:15 PM → 3:45 PM **Networking** 🕒 30m

3:45 PM → 4:45 PM **Forum discussion #1: ESCAPE impact on Scientific Computing in the Research Community.** 🕒 1h

Forum discussion. Convener: Simone

Speaker: Simone Campana (CERN)

4:45 PM → 5:30 PM **Break-out thematic discussions**

7:00 PM → 10:30 PM **DIOS meet-up dinner&more** 🕒 3h 30m

Meeting Agenda: Wednesday

WEDNESDAY, JUNE 8



10:00 AM → 10:45 AM **Forum discussion #2: Rucio and the ESCAPE community, activities, collaboration and synergies** ⌚ 45m

Forum discussion. Convener: Martin, Rob, Rosie

Speaker: Martin Barisits (CERN)

10:45 AM → 11:30 AM **Forum discussion #3: AAI/IAM and token-based auth, future perspectives, Collaborations and fora** ⌚ 45m

Forum discussion. Conveners: Federica, Rizart

Speakers: Federica Agostini (CNAF-INFN), Rizart Dona (CERN), Rohini Joshi (SKA Organisation), Rohini Joshi

11:30 AM → 12:00 PM **Networking** ⌚ 30m

12:00 PM → 12:45 PM **Forum discussion #4: Analysis Platforms, Analysis Facilities. Linking data access and analysis with computing resources.. Scoping future work and identifying interested communities and related fora.** ⌚ 45m

Forum discussion. Conveners: Yan, Elena

Speakers: Elena Gazzarrini (CERN), Yan Grange (ASTRON, the Netherlands Institute for Radio Astronomy)

12:45 PM → 1:30 PM **Forum discussion #5: Consolidating our collaborations** ⌚ 45m

Forum discussion:

- Consolidate our collaborations: summary session outlining goals / directions for collaboration areas
- Register interest of individuals to be on mailing lists and involved in activities
- Share details of related job vacancies coming up

1:30 PM → 3:00 PM **Break-out thematic discussions** ⌚ 1h 30m



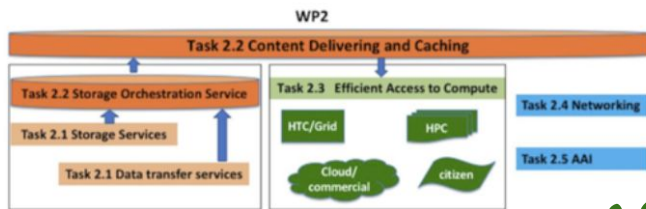
ESCAPE Kick-off Meeting Annecy, February 2019



WP2 tasks

- Task 2.1 Data Lake Infrastructure and Federation Services. CERN (Xavier Espinal)
- Task 2.2 Data Lake orchestration service. DESY (Patrick Fuhrmann)
- Task 2.3 Integration with Compute Services. NOW-I-ASTRON
- Task 2.4 Networking. SKAO (Rosie Bolton)
- Task 2.5 Authentication and Authorization. INFN (Andrea Ceccanti)

Simone Campana (CERN) as WP leader, Rosie Bolton (SKAO) as deputy



ESCAPE Kick-off Meeting

Data Infrastructure for Open Science (DIOS)

- Goal: design, implement and operate a cloud of data services for open access and open science at the Exabyte scale
- The backbone of the Data Lake are well experienced large national data centers supporting the ESFRIs in ESCAPE
- The data lake will serve as underlying data infrastructure to manage and serve data to the user communities
- This solution will be proposed as key component of the future EOSC framework, supporting FAIR principles



ESCAPE
OSSR

Catalogue & Repository of resources

Datasets
Software & services
Tutorials
Training
Publications

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

ESCAPE DIOS Data Lake

FAIR data management
Content discovery and delivery

HPC

PRACE

EuroHPC
Joint Undertaking

HTC

Grid clusters, etc

Private/public clouds

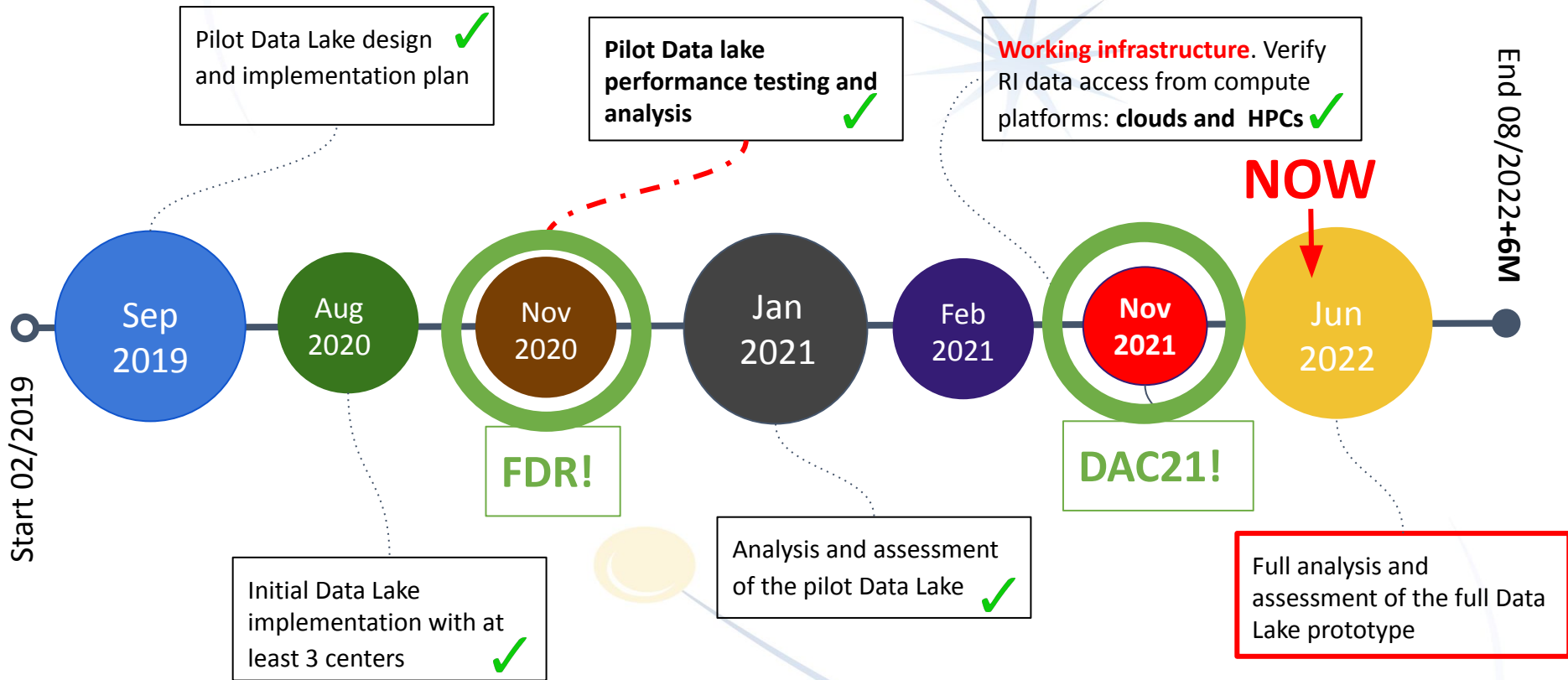
Commercial clouds

GÉANT





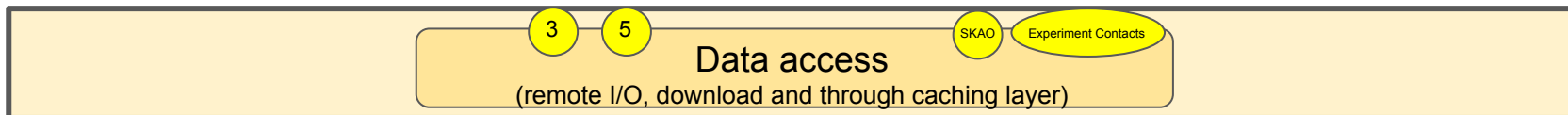
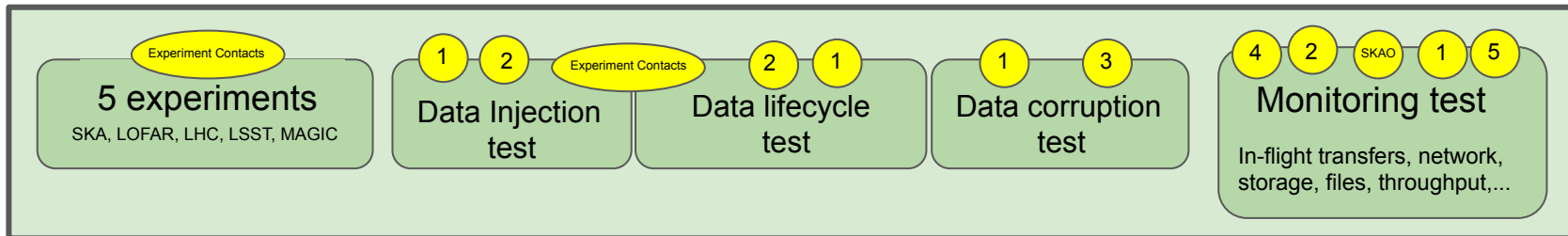
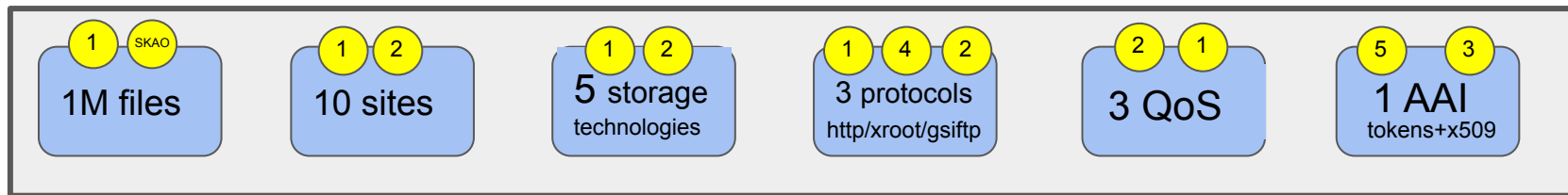
ESCAPE DIOS Roadmap



“24h” pilot datalake performance assessment

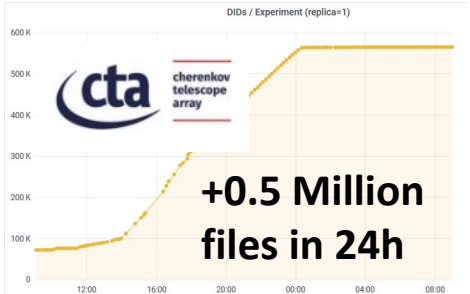
Goal: Exercise covering **experiment data workflow** needs on a single day. From data injection, to data replication and data access. Three fold goal: perspective from **scientists**, perspective from **sites**, and the assessment of the **ESCAPE datalake tools and services** under **pseudo-prod conditions**: RUCIO, FTS, CRIC, IAM, PerfSONAR, monitoring, QoS, clients, etc.

Work plan: Sept-Oct preparations and tests of the different components in order to run together on a single day, ‘a la’ **dress rehearsal**, mid-November (first challenge, probably a 2nd go 15 Dec)

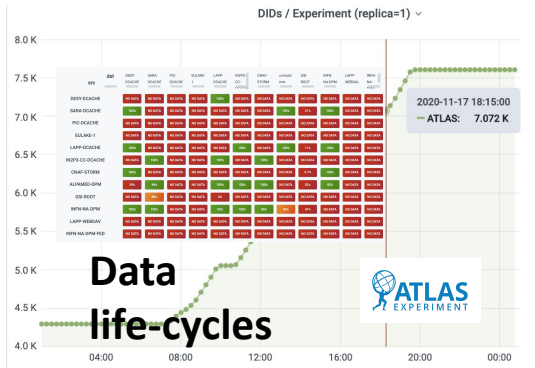


1- datalake	2- QoS	3- integration with compute	4- Networking	5 -AAI
-------------	--------	-----------------------------	---------------	--------

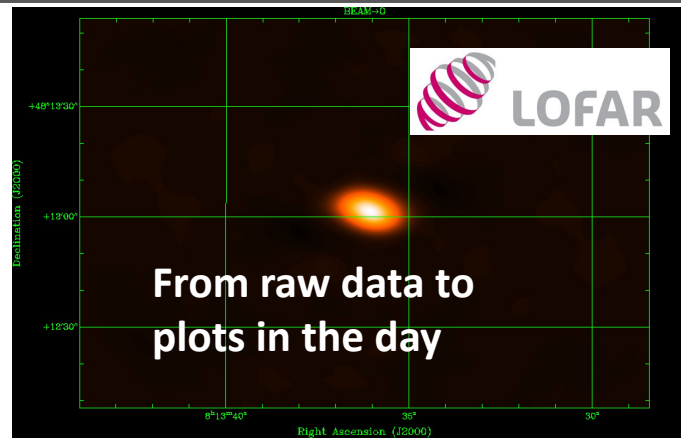
Data Lake 24-hour Dress Rehearsal 17 Nov 2020



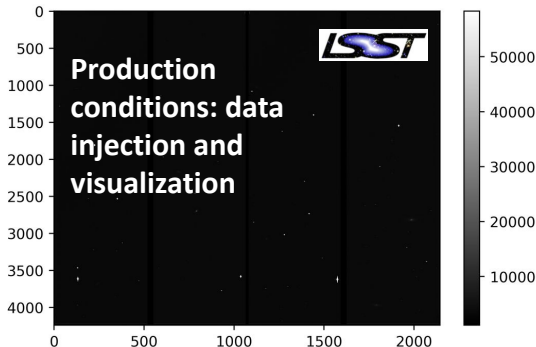
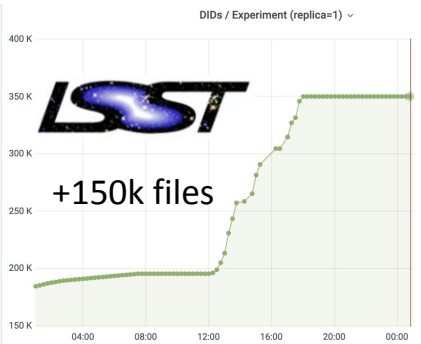
CTA: Simulate a night data captured from telescope in Canary Island for 6 h: ingest 500 Dataset of 10 files.



ATLAS: Storage QoS functionality tests: upload files from LAPP cluster to ALPAMED-DPM (FRANCE) and INFN-NA-DPM (ITALY), then request transfer to 1 RSE **QoS=SAFE** and 2 RSEs **QoS=CHEAP-ANALYSIS**



LOFAR: astronomical radio source 3C196 made using LOFAR data. The raw visibility data was downloaded via rucio from the EULAKE-1 and processed on Open Nebula at surfsara using the container based LOFAR software



LSST: Simulate production conditions: ingest the HSC RC2 dataset from CC-IN2P3 local storage to the Data Lake, **at a realistic LSST data rate (20TB/24h)**. Then **confirm integrity and accessibility of the data via a notebook**.

→ The image is a reconstruction drawn within a Jupyter Notebook accessing the data used in the Full Dress Rehearsal.

From a Data Lake Pilot to a full Prototype (1/3)



WP2 work plan focused on a continuous assessment and evolution of the pilot Data Lake, with the target to meet ESFRI/RI requirements and resulting in a fully working system

- **Token-based authentication:** boosted its integration in the several layers of the Data Lake infrastructure: Rucio, FTS, storages (wip) and integration with other AAI *providers*. Easing user experience with a single and global authentication point
- **Data life-cycle accommodation:** ESFRI/RIs users are able to define data replication rules, lifetimes, access policies, data location and storage *quality of service* (adjusting storage cost with data value)
- **Webdav/HTTP:** promoted to be the de-facto standard in the Data Lake. The widespread knowledge of HTTP protocols provide a flexible way to interact and integrate with other storage resources, also eases data access from heterogeneous compute platforms and end-user devices
- **Rucio consolidation:** two extra ESFRI/RI private Rucio instances in operation for SKA and CTA, harmonically using the same global Data Lake storage infrastructure. KM3Net adopted Rucio after the DAC21!



From a Data Lake Pilot to a full Prototype (2/3)



- **Rucio Evolution:** channeling feedback from the new scientific communities using Rucio. Discussions on extending metadata capabilities together with VO and ESO
- **Enlarged Data Lake monitoring capabilities:** providing real time follow up for data transfers, automated test suite results, resources usage
- **Active Deployment and Operations (DepOps) team:** early in the project identified need to share expertise, organised via a well-established meeting. Crucial to consolidate the infrastructure, to foster knowledge transfer and to prepare and drive the data challenges
- **Expanded Data Lake capabilities with the user environments:** finalised a product labelled as *Data Lake as a Service* (DLaaS),
 - From the notebook it provides to the end users increased data browse/download/**upload** capabilities, trigger data movement, integrate with local storage, leverage storage caches, etc.
 - Allowing to extend functionalities of Analysis Platforms (in conjunction with WP5), and to leverage computing infrastructures (ie. local batch systems and external resource providers)



From a Data Lake Pilot to a full Prototype (3/3)



- **Integration of heterogeneous resources** has been demonstrated, Data Lake interfacing with commercial clouds, public clouds and HPCs
 - **Clouds:** DL flexibility enabled integration of AWS/GCP. Integration verified as 1) computing resource (Monte Carlo generation and data upload) and 2) pure storage node. Also public clouds implemented from partners.
 - **HPC:** CMS PoC with CINECA/HPC, and started a collaboration with the FENIX/HPC project

*The efforts during this second period culminated in a global **Data and Analysis Challenge (Nov-2021)**, certifying the infrastructure as a fully working system*

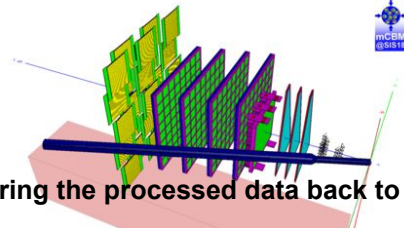
... and brought together scientific communities by joining efforts and ideas, towards common goals in a common infrastructure



Putting the system to work: the DAC21 exercise (1/3)



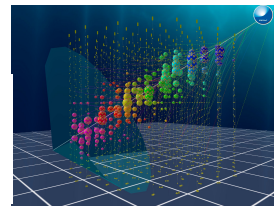
- **Registration of RAW data acquired by the mCBM detector on FAIR-ROOT**
- Ingestion and replication of simulated R3B data
- Ingestion and replication of simulated and digitised raw PANDA fallback data
- Particle-transport and digitisation of Monte-Carlo events
- Live ingestion of simulated data
- **Retrieval of stored RAW data from the data-lake, processing of the data and storing the processed data back to the data lake**
- Retrieve raw mCBM data from the data lake, run reconstruction on it and store the results
- Analyse simulated R3B data stored in the data lake, upload resulting histograms and hitmaps to the DI



Raw data injected, stored and preserved in the DL. Data processed by users, results are stored back in the DL.



- **Ingestion of raw data from the storage at the KM3Net shore station to the Data Lake, and policy-based data replication across the Data Lake infrastructure**



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



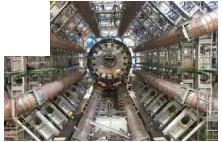






- Long-haul transfer and replication. CTA-RUCIO @PIC: non-deterministic (La Palma) and deterministic (PIC) RSEs
- **Data reprocessing. Primary data stored and findable in the datalake (using the CTA Rucio instance). Data is accessed and processed. New data products stored back in the Data Lake**
- Data analysis. Data access via Jupyterhub/mybinder via ESAP. Higher-level analysis products produced



Distributed data re-processing taken at remote locations

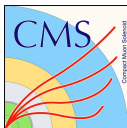
Putting the system to work: the DAC21 exercise (1/3)

 	<ul style="list-style-type: none">• Exercises (data production, replication and documentation) before and during the DAC21. Include the creation of datasets for real-kind final user analysis examples using current open access datasets. ~200*10 – 2000 files uploaded in the Datalake. Two copie• User analysis pipeline to <i>Large experiment demonstrating open data capabilities</i> (http://opendata.atlas.cern/software/). Testing and validating the reading access of the samples via Jupyter rucio extension, and running multiple analysis pipelines.	
 <i>Data management from remote locations</i>	<ul style="list-style-type: none">• Long haul raw data ingestion and replication. Data is successfully transferred from the telescope station and replicated to the Data Lake, file deleted on the telescope storage buffer.• Data transfer monitored. Data can be discovered using the CTA-RUCIO instance. of DL3 file. Validating the reading access of the samples via multiple analysis using gammapy library.	
 <i>Full-cycle scientific data management and data processing</i>	<ul style="list-style-type: none">• Ingestion of LOFAR data from a remote site to the Data Lake. Data transfer and replication into off-site storage, after successful replication delete data at the source• Process data in the Data Lake at an external location, combine results with other astronomical data to produce a multiwavelength image.• Include a read-only RSE to a location outside the data lake. Get data from there into the DL.• Extending use cases by using larger files and leveraging several QoS, running all processing in the DLaaS, requiring the availability of specific LOFAR software in the DLaaS.	

Putting the system to work: the DAC21 exercise (1/3)



- Data replication. Data in correct place in timely manner.
- **Long haul data replication. SKAO Rucio (Australia and South-Africa to UK RSEs), using the RUCIO SKA instastream**
- End-to-end **Global-scale Data Management** pipeline from SKA to northern hemisphere sites
- Data analysis. Subsequently running SKA science data challenge pipeline using data stored in datalake.



- **Multi purpose Analysis Facility PoC with data access via DASK (workload orchestrator) leveraging computing at Marconi (HPC) and large batch clusters**
- Access control for embargo data, test in CNAF and DESY
- Content delivery and caching: XCache Protocol Translation: xroot internal vs http External for Data Lake data transfer. Performance comparisons for Analysis workflows



DL interface with local and heterogeneous resources, CDN and caching



- **Simulate replication of one night's worth of raw images data between two Vera C. Rubin data facilities, perform the exercise several times.** Each iteration is composed of 15TB, 800k files, ideally to be replicated in 12 hours or less
- Incorporate SLAC National Accelerator Laboratory (US) in the data replication chain (postponed)



Leverage telescope local storage data replication to fulfill daily data management cycles

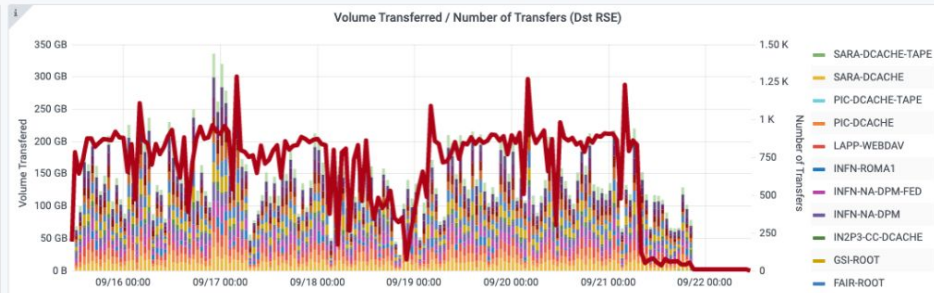
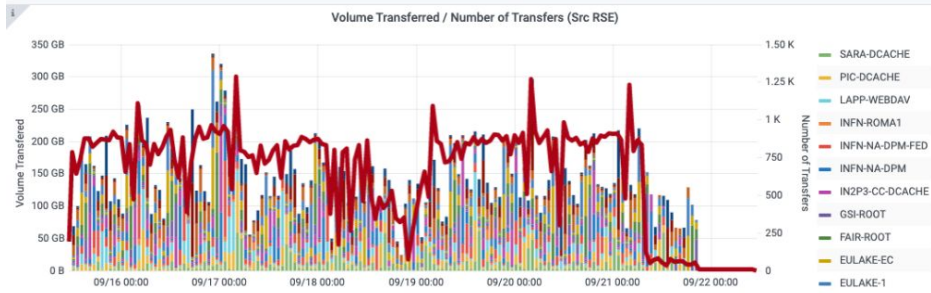
E-EAB Recommendations



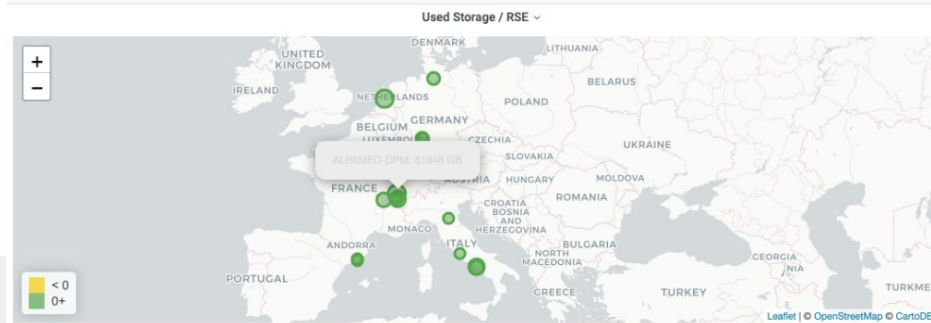
E-EAB first review recommendations

- WP2 Data Infrastructure for Open Science (DIOS):** D2.1 Implementation plan and design of pilot presents excellent progress in cataloguing the available tools and systems, and the identification of the best options for the prototype-phase is to be started now. It would be useful to see how the data-lake connects more concretely with the various ESFRIs and pan-European research infrastructures, by locating their existing storage/computing/services on the data-lake architecture diagram (or a separate one). The implementation plan should be completed with the corresponding risk analysis.

Volume Statistics



World Map Panels



Feedback from the second review (1/3)

REC1: DIOS should provide details and concrete examples on how ESFRI data currently in their science archives can efficiently be stored and accessed through the ESCAPE data lake.

The current storage used in the experiments can be attached to the Data Lake by providing the storage endpoint and specifying the storage system protocol.

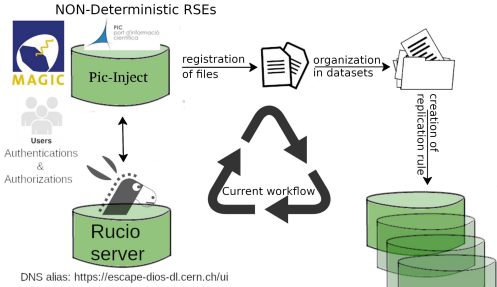
The registration of files already present in the Storage Element is done by scanning the namespace and populating Rucio Catalogue entries. This is a pure metadata operation with file checksum verification.

Approach validated: the MAGIC Telescope is using a local storage sitting next to the telescope, this storage was made Data Lake aware in order to control the file replication from La Palma island to the Data Lake, and also handling data deletion once the files are well consolidated the DIOS to free up buffer space in the telescope storage.

Future use case addressing a real need: LSST (Vera C. Rubin Observatory) telescope sits in the Atacama desert, data is sent to SLAC and then replicated to Europe with N2P3CC as entry point.

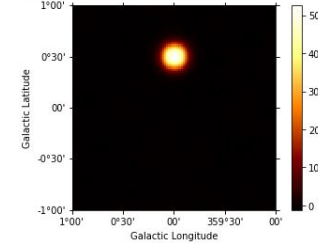


Example-1: Full-cycle long-range data workflows



```
[38]: # we can also compute the significance of our source
analysis.get_excess_map()
analysis.excess_map["sqr_ts"].plot(add_cbar=True);
```

Computing excess maps.
Position <SkyCoord (Galactic): (l, b) in deg
(0., 0.)> is outside valid IRF map range, using nearest IRF defined within

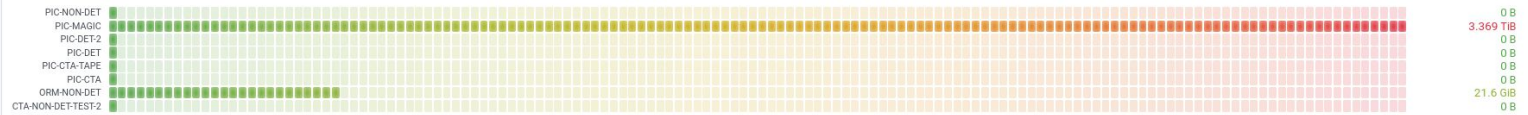
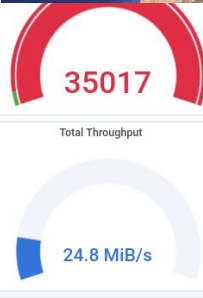
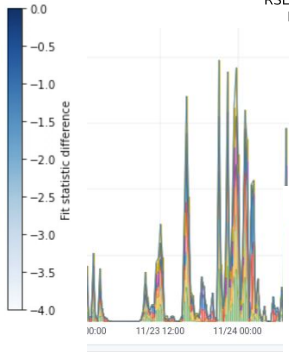
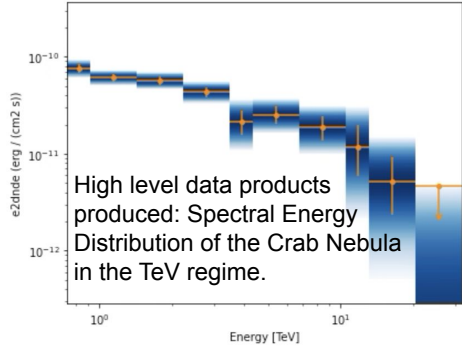


perform the fit

As a final step we fit the spectrum of the source, and we compare to the one we actually used for simulation

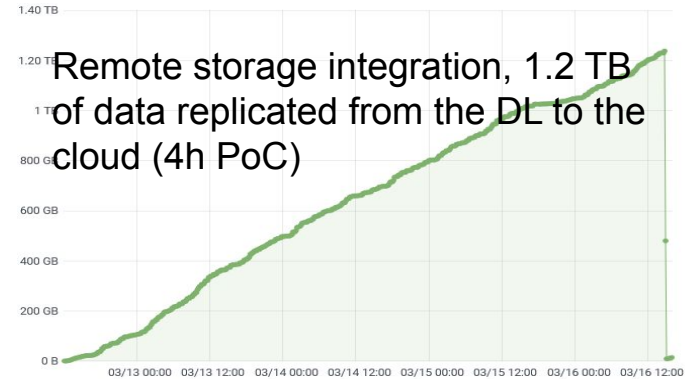
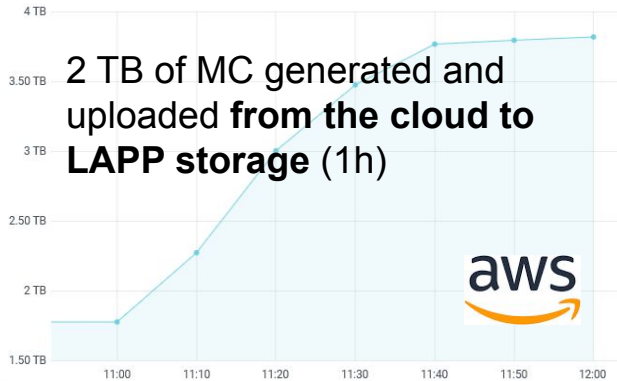
```
[43]: # let us load the model we used for the simulation
models = Models.read("./data/models/point-source-pwl.yaml")
# let us create a copy of the spectral model for later comparison
original_spectral_model = models[0].spectral_model.copy()
```

Source	Destination	V0	Submitted	Active
+ gsiftp://datatransfer.ctan.cta-observatory.org	gsiftp://door05.pic.es	pic01-rucio-server.pic	5589	129
			5589	129

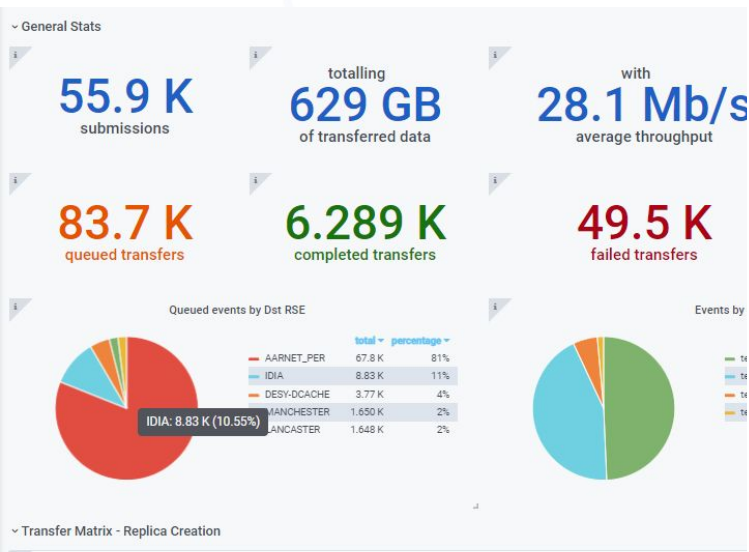


Example-2: Integration with commercial and public clouds

- Goal: assess integration of heterogeneous resources within the ESCAPE DL, Including CPU and storage using industry standards (Swift/S3 protocol)
- Exercise performed with the support of the [Cloud Bank EU NGI](#) project with fundings for AWS and Google Cloud Platform
- *Use case 1: Generation of CTA's Monte Carlo and results upload to the Data Lake*
- *Use case 2: Ad-hoc integration of Commercial Cloud storage in the Data Lake*



Example-3: Extreme distance data management in SKAO



Storage Endpoints

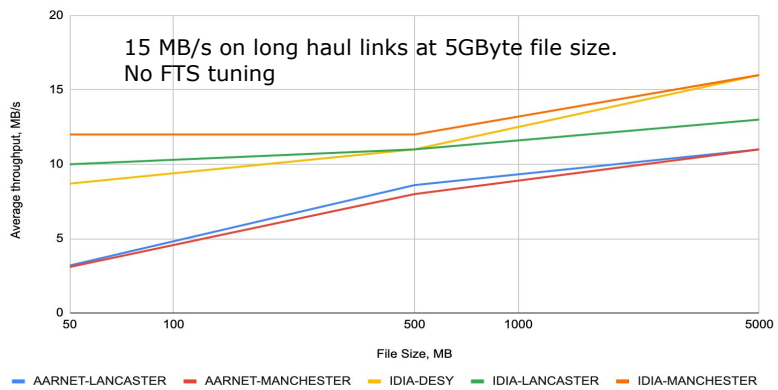
European, South African and Australian sites
Best efforts basis currently

Mix of deterministic and non-deterministic RSEs (**non-deterministic to mimic SKA data staging source storage**)

Instance will continue to grow over coming year, as SKA Regional Centre partners join and help assess functionality. Anticipate Spain, China, France, and maybe Canada



Long-haul transfers from Australia and South Africa to European locations during DAC21



Manual data transfers with Rucio began Feb 18th 2021 and automated tests running hourly across all sites since Feb 23rd

Feedback from the second review (2/3)

REC2: DIOS and CEVO WP should work more closely together to demonstrate full interoperability between the ESCAPE data lake and data / metadata storage and access through VO protocols.

The synergies between CEVO and DIOS took some time to build up. This was expected. But as both work packages were evolving the gap between them was shortening. We are in a position where the common interests have been identified and synergies start flowing on the technical side and also from the conceptual perspective.

A concrete example of Data Lake and VO files/data integration is with the HIPS catalogue, where the possibility to have the DIOS as the backend for the HIPS files is being assessed.

On the other hand, the large metadata requirements by the IVOA community and in particular ESO is a good stepping stone to start bridging the different approach on Data and Metadata catalogues between Physics and Astronomy *(see later analysis done merging radio-astronomy data with visible spectrum image)*

The goal involving DIOS/CEVO for this final phase of the project is to identify the synergies between the communities and set the right collaboration channels. Start focusing on an identified subject and build-up, ie. HIPS and DIOS, extension of Rucio metadata capabilities.



Example: multi-wavelength analysis



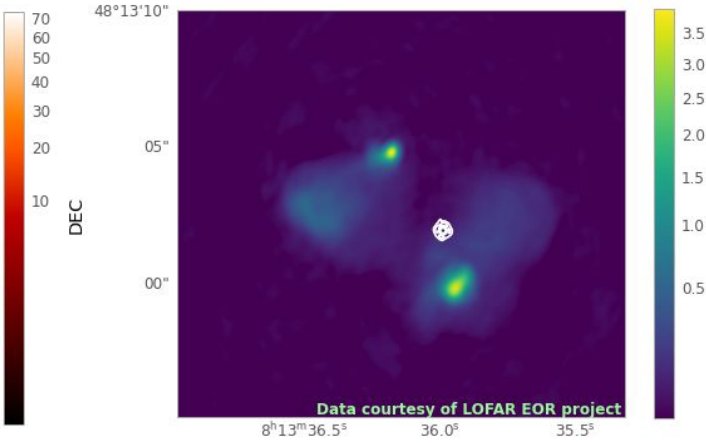
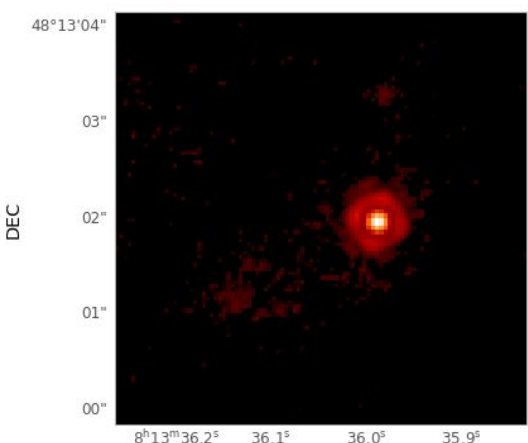
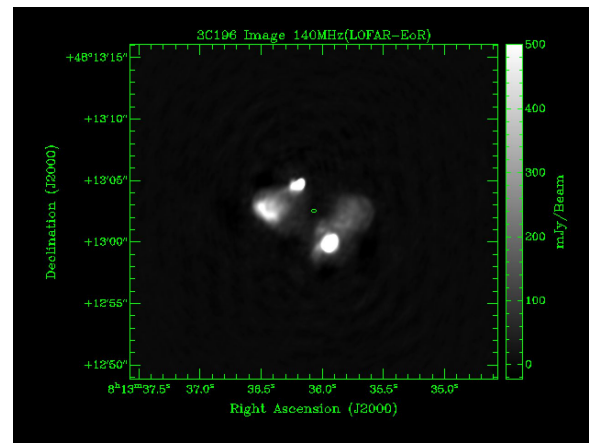
1. Data injected to the DL from **three** radio source observations in external locations

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

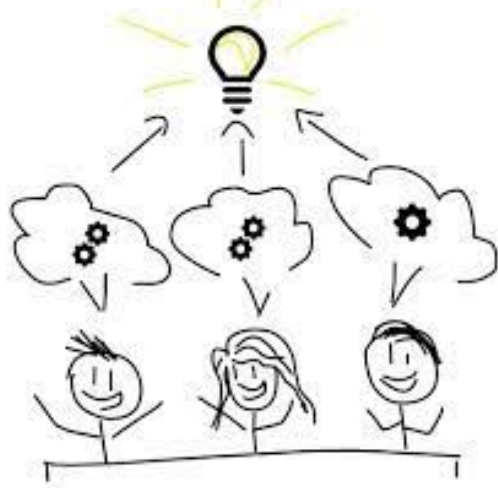
4. Combined optical data from the Hubble located via the **VO (WP4)**

5. Optical and radio data aggregate in via the **ESAP (WP5), combined analysis done**. Results uploaded back to the DL.

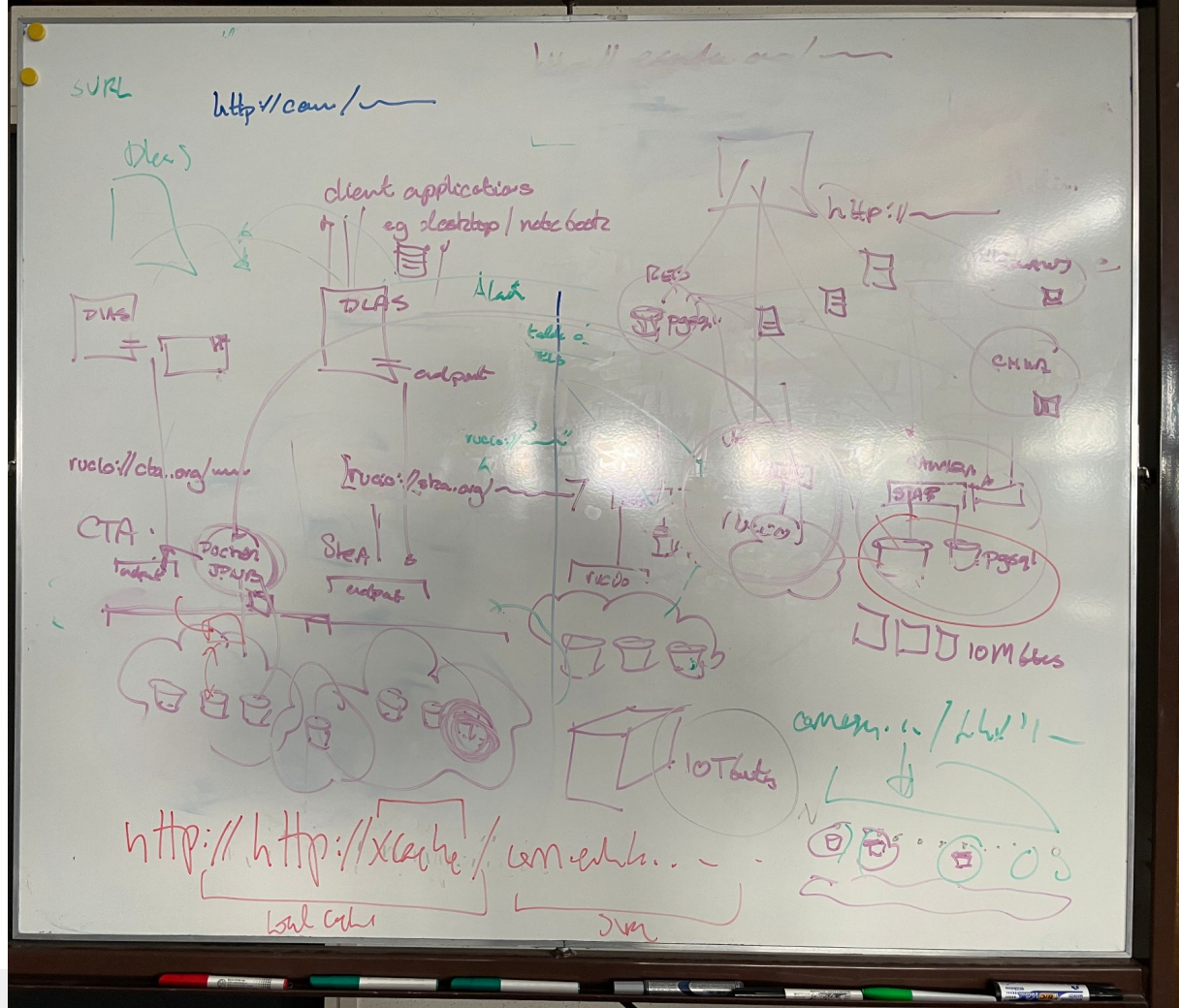


From left to right: Radio image, Optical image and the Combined image (LOFAR with optical contours)

From the 24/ 06/22 extended EEB!



2h whiteboard session



Feedback from the second review (3/3)

REC3: DIOS should list the data required to implement the 2 TSPs, where they will be stored and how they will be accessed.

The data required for the TSPs will come from the Open Data Portal and/or injected on demand by the Postdocs working in the RIs.

The data will be stored in the ESCAPE Data Lake, making sure the right level of replication is achieved and enforcing geographical availability to minimise data access latencies.

The plan to access the data is to exploit the advantages of the Virtual Research Environment (VRE), offering remote data access (file is read via the network), local file replication to the computing node (notebook backend) or via a Content Delivery Network based on data caching technologies maximising file reusability and minimising access latency, as the data caches are close to the computing nodes by design.



Example: ATLAS Dark Matter Reinterpretation - Dilepton Resonance

Import files from the Data Lake into the notebook

RUCIO

EXPLORE NOTEBOOK

ATTACHED DIDS

- ATLAS_LAPP_SP:DMCrossSectionGraphs_axial_ee.root axial_ee
- ATLAS_LAPP_SP:DMCrossSectionGraphs_axial_mumu.root axial_mumu
- ATLAS_LAPP_SP:LimitInterpolator_CL95_14TeV.root limit_intepol

```
[12]: axial_ee, axial_mumu, limit_intepol

[12]: (/eos/eulake_1/ATLAS_LAPP_SP/9d/f2/DMCrossSectionGraphs_axial_ee.root,
/eos/eulake_1/ATLAS_LAPP_SP/58/50/DMCrossSectionGraphs_axial_mumu.root,
/eos/eulake_1/ATLAS_LAPP_SP/23/c7/LimitInterpolator_CL95_14TeV.root)

[9]: import ROOT
import gfal2

[13]: type(axial_ee)

[13]: rucio_jupyterlab.kernels.ipython.types.SingleItemDID

[11]: def GetInteg(histo):
return histo.Integral()

def getDMCrossSection(medType):
outfilename = "DMCrossSectionGraphs_ " + medType
```

Final results

```
if finalState == "ee": leg.AddEntry(expLimit, "#font[42]{Expected e^{+}e^{-} limit}", "l")
else: leg.AddEntry(expLimit, "#font[42]{Expected #mu^{+}#mu^{-} limit}", "l")
leg.AddEntry(fidXsec, "#font[42]{Vector Z'_{DM} (m_{chi1}=#mDM+ TeV)", "l")
leg.Draw()
ROOT.gPad.RedrawAxis()

fOutput.cd()
if mDM == "0.50": expLimit.Write()
fidXsec.Write()
c.Write()
c.SaveAs("dilepton_jared/output/Crossing_DM"+massDM+"_fs"+finalState+".pdf")

return expLimit, fidXsec

def DrawAllCrossing(fOutput, finalState):
massDM = ['0p50', '1p00', '1p50', '2p00']
for mDM in massDM:
MakeCrossing(fOutput, finalState, mDM)

if __name__ == "__main__":
ROOT.gROOT.SetBatch(True)
```

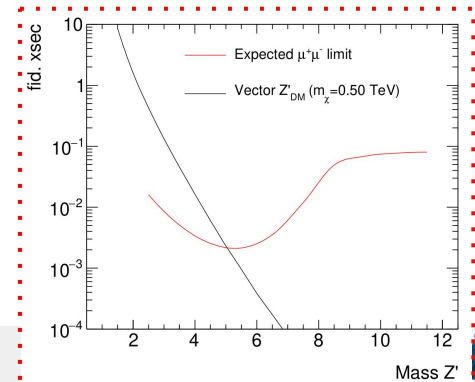
Analysis Preservation: Ensure analysis reusability and reproducibility (REANA)

```
1 version: 0.8.1
2 inputs:
3   directories:
4     /python/
5     /data/
6   files:
7     /python/MakeLimit.py
8     /python/Summary.py
9     /data/DMCrossSectionGraphs_axial_massass.
10    /data/LimitInterpolator_CL95_14TeV.root
11 workflows:
12   type: serial
13   specification:
14     steps:
15     - name: SetLimits
16       environment: 'reanahub/reana-env-root6:
17         compute backend: kubernetes
18         kubernetes: 'limit: '9Gi'
19     - name: compute
20       environment: 'python/python:MakeLimit.py
21     - name: plots
22     - name: plots
23     - name: plots
24     - name: plots
25     - name: plots
26     - name: plots
27     - name: plots

jovyan@jupyter-egazzarr:~/dilepton_jared/atlas-dm-reinterpretation$ ls
notebooks python README.md reana.yaml runReana.sh
jovyan@jupyter-egazzarr:~/dilepton_jared/atlas-dm-reinterpretation$ ls
notebooks python README.md reana.yaml runReana.sh
jovyan@jupyter-egazzarr:~/dilepton_jared/atlas-dm-reinterpretation$ bash runReana.sh
==> Verifying REANA specification file... /home/jovyan/dilepton_jared/atlas-dm-reinterpreta
-> SUCCESS: Valid REANA specification file.
==> Verifying REANA specification parameters...
-> SUCCESS: REANA specification parameters appear valid.
==> Verifying workflow parameters and commands...
-> SUCCESS: Workflow parameters and commands appear valid.
==> Verifying dangerous workflow operations...
-> SUCCESS: Workflow operations appear valid.
==> Verifying compute backends in REANA specification file...
-> SUCCESS: Workflow compute backends appear to be valid.
SettingLimits.1
==> SUCCESS: File /python/MakeLimit.py was successfully uploaded.
==> SUCCESS: File /python/Summary.py was successfully uploaded.
```

Also HTCondor and Slurm

30



ESCAPE, so far...



- Provided a framework to **explore** and **influence** on the development of next generation distributed computing models and data management tools
- Achieved a **fully working system**, ahead of the originally planned prototype
- Key of the success? your **engagement**
- Set the scene for further projects and collaborations with the right spirit: sites, ESFRI/RIs and service providers collaborating, working and exchanging **together**
- Sheer interest from external communities and projects to follow up the work being carried in ESCAPE, huge synergies, huge opportunities, **exciting future!**



ESCAPE is ending... long life to ESCAPE



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”**.

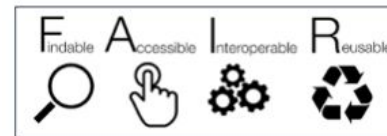
Other Science Clusters: **ENVRI-FAIR** (Environment and Earth Sciences), **EOSC-LIFE** (Biomedical Science), **PANOSC** (Neutron and light sources facilities) and **SSHOC** (Social Science and Humanities).



Five Science Clusters



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



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

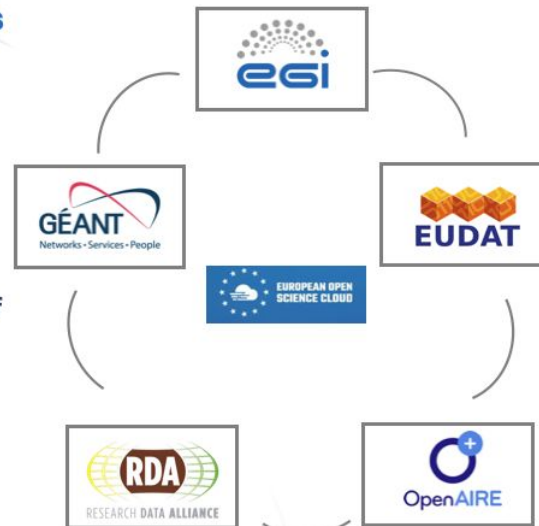
**The ESCAPE Virtual Research Environment prototype
to host the Test Science Projects
(and part of the EOSC Future project [...])**



TSPs and synergies in EOSC Future

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

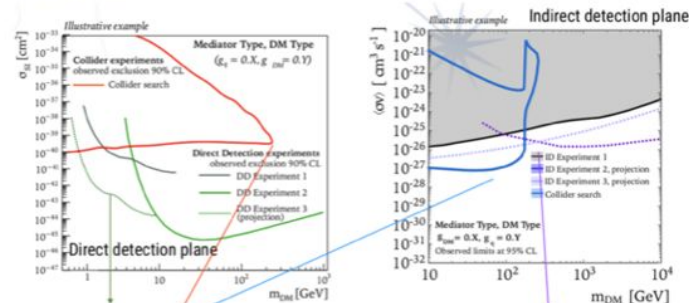


Open Science Pilots within “Dark Matter” as focus / scientific question

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

Slides by E. Gazzarrini and DM Science Project organizers (F. Calore, C. Doglioni, L. Heinrich)

Funded by the European Union's
Horizon 2020 - Grant N° 824064



A series of pilots focused on violent phenomena in the Universe with **Astrophysical** 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



GENERAL PROJECT REVIEW CONSOLIDATED REPORT

Grant agreement (GA) number:	824064
Project' Acronym:	ESCAPE
Project title:	European Science Cluster of Astronomy & Particle physics ESFRI research infrastructures
Type of action:	RIA
Start date of the project:	01/02/2019
Duration of the project:	48
Name of primary coordinator contact and organisation:	Giovanni Lamanna (CNRS)
Period covered by the report:	from 01/08/2020 to 31/01/2022
Periodic report/Reporting period number:	2
Date of first submission of the periodic report (if applicable):	22/02/2022
Amendments (latest AMD concerning description of the action) ²	22/09/2021 (AMD-824064-32)
Date of meeting with consortium (if applicable):	02/03/2022
Name of project officer:	Christos CHATZIMICHAIL
Name(s) of monitors:	<ul style="list-style-type: none"> - Anne Claire Mireille FOUILLLOUX <ul style="list-style-type: none"> • University of Oslo • ECMWF (European Centre for Medium-Range Weather Forecasts) • IDRIS(Institut du développement et des ressources en informatique scientifique) - Jan HRUSAK <ul style="list-style-type: none"> • Academy of Sciences of the Czech Republic • J. Heyrovsky Inst. Physical Chemistry • Ministry Education Youths and Sports - Natalia BELOFF <ul style="list-style-type: none"> • University of Sussex

29/9/2021

As from the last ESCAPE-RP2 review report of EC:

"...The role of ESCAPE in the construction of EOSC is key but could also be increased by taking the initiative to explain its approach to other ESFRI cluster projects..."

"...It is important for other scientific communities to "apply" these (ESCAPE) technologies to" other use cases": this would help to engage with more users and increase the benefits of EOSC..."

Recommendations:

« ... promote ESCAPE strategic conceptual and unified vision of EOSC outside the ESCAPE community, including other ESFRI cluster projects. This would also benefit the ESCAPE community, through new collaborations/projects (for instance, exploitation of services related to the ESCAPE data lake, citizen projects and engagement with users), recognition of the ESCAPE community as a leading Open Science and actively contributing to the construction of EOSC...

.. ESCAPE has the capacity to be a natural leader of the other RI related projects when it comes to the contribution to the EOSC concept evolution and the implementation of concrete EOSC services. These services must be developed according to the user needs and it is recommended for ESCAPE to continue integrating the knowledge that was accumulated by the RIs and piloted in interactions with user communities and the broader society...

... The ESCAPE project is very well positioned to be an amplifier of the voice of data producers and data curators in EOSC and Open Science community. »



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/2432> 



Preliminary and confidential perspectives – tentative list of topics

Destination INFRAEOSC : Enabling an operational, open and FAIR EOSC ecosystem

(i) HORIZON-INFRA-2023-EOSC: (CALL 1):

Build on the science cluster approach to ensure the uptake of EOSC by research infrastructures and research communities

- This topic aims to extend the level of cross-domain collaboration and EOSC alignment initiated in Horizon 2020 with the science cluster projects. It also capitalises on the experience gained by these cluster projects in enabling open science practices, FAIR implementation and managing open calls for disciplinary and multi-disciplinary science projects to involve smaller or less structured communities with less experience in open science, and to support communities lacking relevant competence centres.
- It covers both Consolidation and new Science Projects as prospected by GL (in his presentation at the June 2021 SCLs workshop*):
 - ✓ EC expects the five clusters to work together in this CALL 1 (namely, a limited number of partners representing the five clusters ...The new ESCAPE collaboration agreement helps in this sense).
 - ✓ The activity must be inclusive (more world-class RIs, Universities) and shall be implemented through **open calls** for cross-RI and/or cross-domain science projects and services through a **cascading grant mechanism**.
 - ✓ The cluster coordinators support GL naturally to coordinate CALL1. GL is considering to give continuity to his engagement for the 5 SCLs.

(ii) HORIZON-INFRA-2023-EOSC: (CALL 2):

Development of community-based approaches for ensuring and improving the quality of scientific software and code

A framework of community curation is established and promoted that ensures quality of software and code across the different disciplines.

... Develop or align pre-existing training materials for software development skills, digital badges, etc.

- GL: CALL 2 will help consolidating the ESCAPE scientific software catalog, the current innovative developments (AI, Quantum Computing etc.), and will bring closer HSF and Astro communities.



New ESCAPE Collaboration Agreement

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.



ANNEXE 3 – ORGANISATION AND COMMON DUTIES

The Parties commit to set up and support the most appropriate, lightweight and efficient management organisation of the ESCAPE Collaboration. We note that the individual national institutions will be represented through the RIs.

The Parties will set up and appoint members of the following structures:

1. A Director of the Collaboration, elected by the RIs in the Collaboration (1 vote each), with a term of 2 years, that can be re-elected.

2. A Strategy Board (SB)

- a. Members of the Strategy Board are the nominated representatives and alternates (1 representative + 1 alternate) of each of the RI's that are ESCAPE partners.
- b. The SB will be chaired by the Director of the Collaboration.
- c. The function of the SB is to define strategy, agree resources as needed for collaborative projects, coordinate with other Science Clusters, coordinate with the EC, coordinate and collaborate with ESFRI, ERA, and JENAA (and its constituent consortia) and organise response to potential funding calls.
- d. The SB will oversee and ratify the work programme.
- e. The SB will meet 3-4 times per year, or as needed.
- f. The SB can instantiate working groups to address specific issues of policy, strategy, etc.

Collaboration Agreement

3. An Executive Board (EB)

- a. The EB reports to the SB.
- b. The members of the EB are the technical coordinators (or equivalent) of the RIs in the collaboration and leaders of implementation working groups. The chairperson will be nominated by the members. The chairperson will become the Technical Coordinator of the collaboration, and will have a term of 2 years, renewable.
- c. The role of the EB is to:
 - i. Propose to the SB and coordinate agreed technical collaborative projects between the RIs;
 - This can include but is not limited to work on common software, infrastructure, services, etc.
 - The work will be executed through setting up and overseeing working groups with members drawn from the RI's as needed, and leaders of any eventual work package structure who would also be members of the EB.
 - ii. Technical coordination with the EOSC and EOSC-related projects.
- d. The EB will meet monthly or as required.

4. An External Advisory Board (EAB)

- a. Members of the EAB are nominated by the Director and are asked to provide independent advice, to conduct an independent assessment of the progress being made by the ESCAPE collaboration as well as to support connection with the national thematic institutes and the scientific community at large³.

³ The current H2020 ESCAPE EAB is composed of an ESA representative as well as the chairs of APPEC, ASTRONET, ECFA and NuPECC.



ANNEXE 2 – AREAS OF COLLABORATION / WORK PLAN

The initial set of topics of common interest include, but is not limited to:

- 1) Collaboration on common infrastructure and tools, recognising that our National and Institutional data and computing centres support many RIs, and that common services and infrastructure are essential to sustainable and cost-effective operation and support of the scientific communities.
- 2) Continued development of a federated data management infrastructure for FAIR and Open Science, which develops the key features of scalability to multi-Exabytes, reliability, policy-driven replication, and content delivery to distributed processing resources. The data infrastructure should be usable by all the participating RIs and integrated with the European Open Science Cloud core services. Specific services will be proposed for inclusion in the EOSC Exchange layer.
- 3) Develop the repository and catalogue of scientific software, tools and services developed in the ESCAPE project, and continue to integrate into the EOSC portal and EOSC Exchange as appropriate.
- 4) Support and develop the common baseline of the Virtual Research Environments developed in ESCAPE. This should include support for long term research infrastructures such as IVOA and WLGC where appropriate and their inclusion in EOSC.
- 5) Develop a sustainable model of collaborative operations and support for these environments, services and tools. This is critical and must be in place to ensure the support of tools and services developed in ESCAPE and brought into production as key elements of the RI computing environments (for example Data Lake, OSSR, etc). Investigate how some of the operations could be supported by EOSC.
- 6) Collaboration on Citizen Science through continued development and evolution of citizen science infrastructures, tools, and credit systems. This should be integrated with the ESCAPE baseline services as far as reasonable in order to make open research data accessible. Development of open science portal(s) to publish and make available open data sets and tools to exploit them, and mechanisms to make resources available to support such projects. Ideally this would be in collaboration with other Science Clusters and part of a larger commitment to "Science for Society".
- 7) Collaboration on advanced technologies, such as AI/ML, and Quantum computing, useful in support of the data analysis in the ESCAPE scientific domains. These should be driven through strong science use cases, and bring in novel algorithm development and support.
- 8) Collaborate with the HPC community, through the FENIX, PRACE, and EuroHPC partnerships, to ensure that ESCAPE RI's can integrate the use of HPC services as relevant components of an overall computing environment.

- 9) Develop, in collaboration with other Science Clusters, a European Virtual Software Institute for Research software. The concept is to tap into the research knowledge of University CS departments, software engineering schools for the benefit of (natural) science developments. The aim is to:
 - a. Enable R&D resulting from collaborations of CS and natural science
 - b. Establish a career path for scientists and engineers working in software and computing in natural science [in many fields the recognition of software work and finding / retaining experts is a major concern].
 - c. Cross-fertilize knowledge between different science domains and make the acquired knowledge available across domain boundaries.
 - d. Act as a lobbying organisation and raise awareness of software and computing in natural science.

The action should build on work in national RSE projects, software carpentries, etc. including collaboration with SMEs to benefit from expertise, projects, placements, etc.

Organise also a scientific computing conference series, for the broad scientific research community, in collaboration with other Science Clusters.

- 10) Career development for young scientists and training in Astronomy, Astroparticle, astrophysics, cosmology, high energy, and nuclear physics, specialising in scientific computing and research software (this action should be in collaboration with other clusters). This will build upon the activities of a Virtual Software Institute (as in 9).
- 11) Pursue the aims of transversal/multi-domain (Test) Science Projects as in the H2020 EOSC Future project, targeting a second phase within Horizon Europe work programme to uptake new emerging and challenging "Open Science Objectives". Extend commitments from more RIs in current and new Open Science Projects (OSP). ESCAPE will also leverage the inter-cluster coordination for Cross-Cluster Open Science Projects (COSP) and when relevant will act to reach out and support "the long tail" of science and multiple scientific communities.
- 12) Support the European Strategy for Data by exploring and building synergies on "Sector Data Spaces" in which interests of ESCAPE Parties would emerge and for the provision of secure and FAIR-enabling European cloud services.
 For example:
 - a. Cross-sector sharing of data and Green Deal data for a unifying, forward-looking approach of any Big Science facility for energy efficiency, water management, support of circular economy and any environmental implications of RIs construction, etc.
 - b. Health data linked with high energy particle and nuclear physics facilities for preventing/treating diseases.
 - c. Industrial and Manufacturing data by exploiting FAIR digital objects from the R&D programmes for detectors, sensors, telescopes and other devices of the ESCAPE RIs.
 - d. Opening data and innovation projects to training actions by research for a Skills data space, to reduce the skills mismatches between the education and training systems and the labour market needs.

It is understood that further technical and policy topics may be added to the above list.



- Why have ESCAPE partner institutes not been involved/informed about the CA? Why are institutes not being asked to be partners?
 - *Institutes will be included via the RI's (but should still be informed now after having received the RI's approval)*
 - ...
- How will new experiments or RI's join?
 - Envisage associate members - can collaborate technically, but not define strategy or vote; eventually associate members could become full members.
 - ...
- How will the ESCAPE collaboration join new HE calls?
 - Individual partners will join projects to represent ESCAPE, and coordinate between project and ESCAPE (like we do in EOSC-Future)
 - some other partners might be linked 3rd parties
 - ...
- How will effort be funded to execute the proposed work programme?
- (or ... Why have you produced a work programme if there is no funding?)
 - RI's will work on / collaborate on topics that they need - presumably it satisfies a requirement that they have anyway; but collaborative software/services are stronger/more sustainable than going it alone - we should not have a work programme of topics that would not be useful to the RI's
 - via Horizon Europe projects - for some relevant partners; this could add additional tasks that we may not have otherwise done as a "cost" of the project funding



Enjoy the meeting!

