ESAP: ESFRI Science Analysis Platform (WP5)

John Swinbank — swinbank@astron.nl

WP5 / ESAP in the ESCAPE "big picture"

WP5 / ESAP in the ESCAPE "big picture"

TSP s

RI-Specific Science Platforms

ESCAPE ESAP Science Platforms

Workflows, notebooks, deployment platforms, packaging
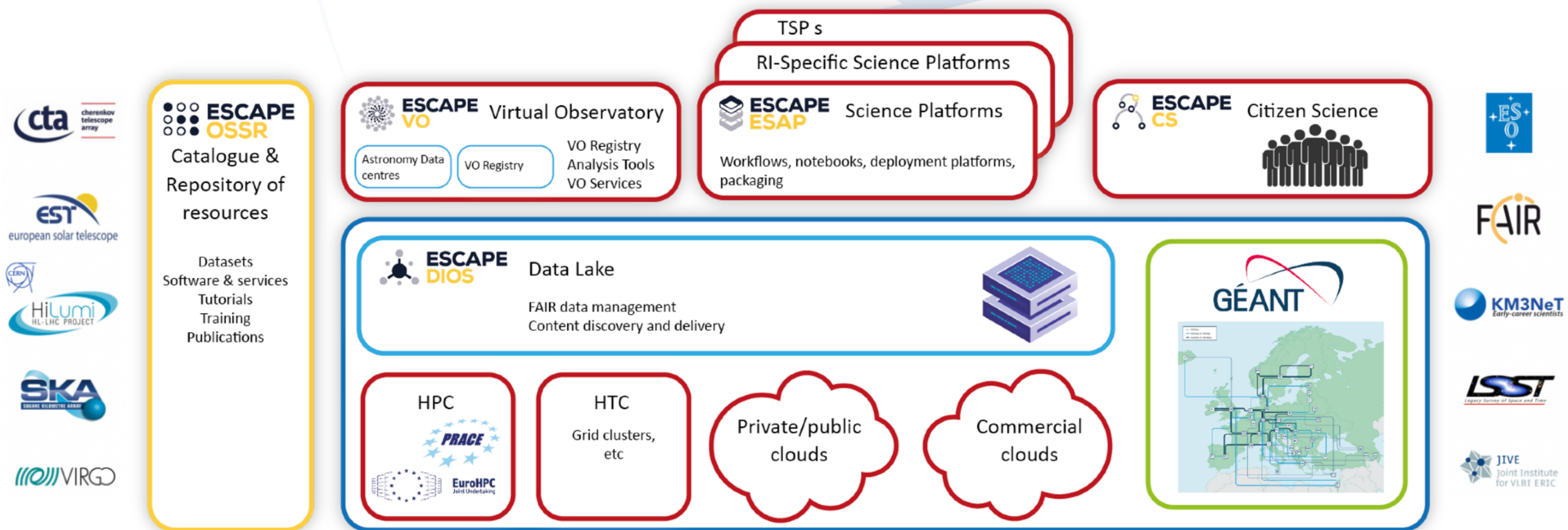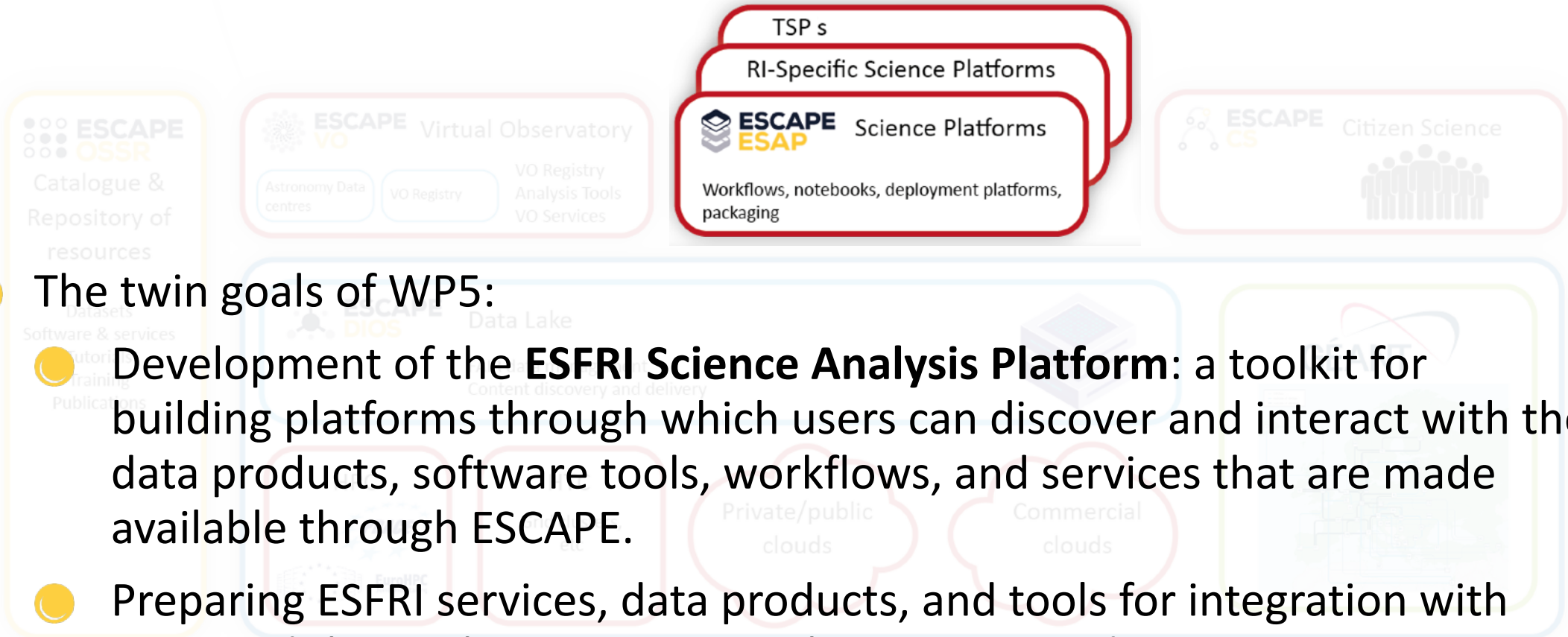
- The twin goals of WP5:
  - Development of the **ESFRI Science Analysis Platform**: a toolkit for building platforms through which users can discover and interact with the data products, software tools, workflows, and services that are made available through ESCAPE.
  - Preparing ESFRI services, data products, and tools for integration with ESAP and their subsequent use within ESCAPE and EOSC.
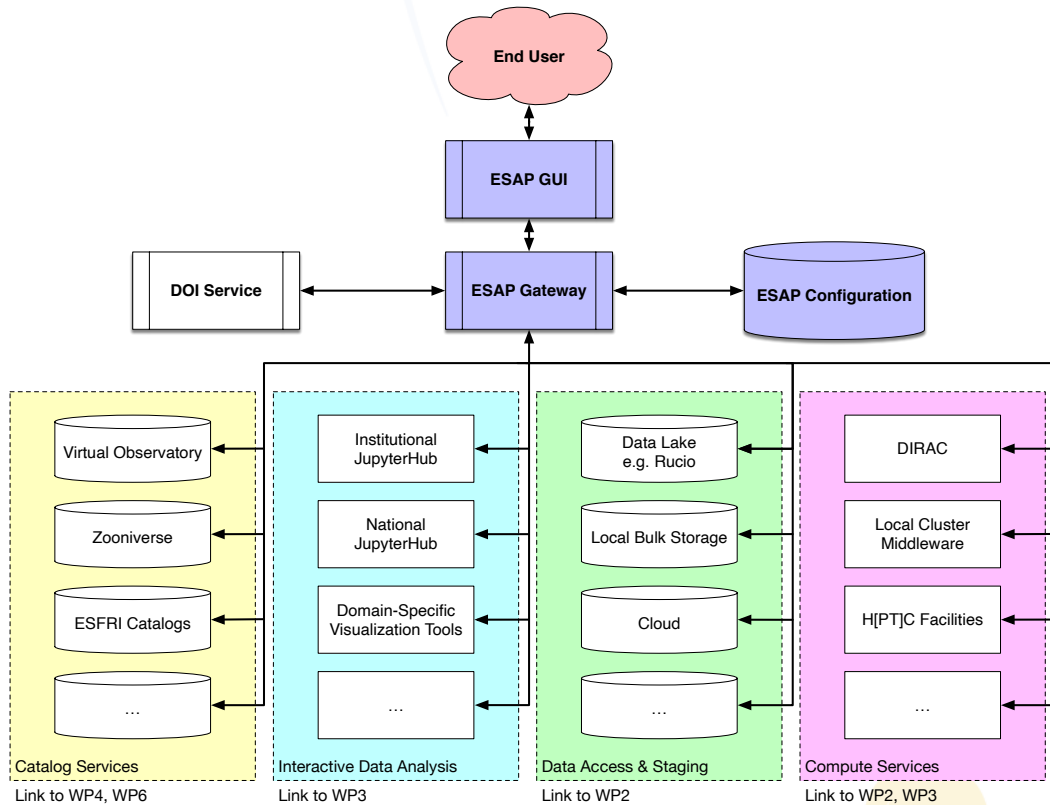
ESAP deliverables

- ESAP provides a focal point for integrating diverse services which are drawn from other providers.

- The heart of the system is the web-based *User Interface* and the *API Gateway*. These software products are the core WP5 deliverables.

- The API Gateway modulates communication with a range of external services using REST APIs. A "service connector" system provides a standardized, plugin-based mechanism for integrating new services.

- WP5 members and ESFRI partners collaborate on developing & integrating a variety of external services.

- It is possible to extend ESAP to operate in a variety of different environments by implementing & integrating new service connectors.
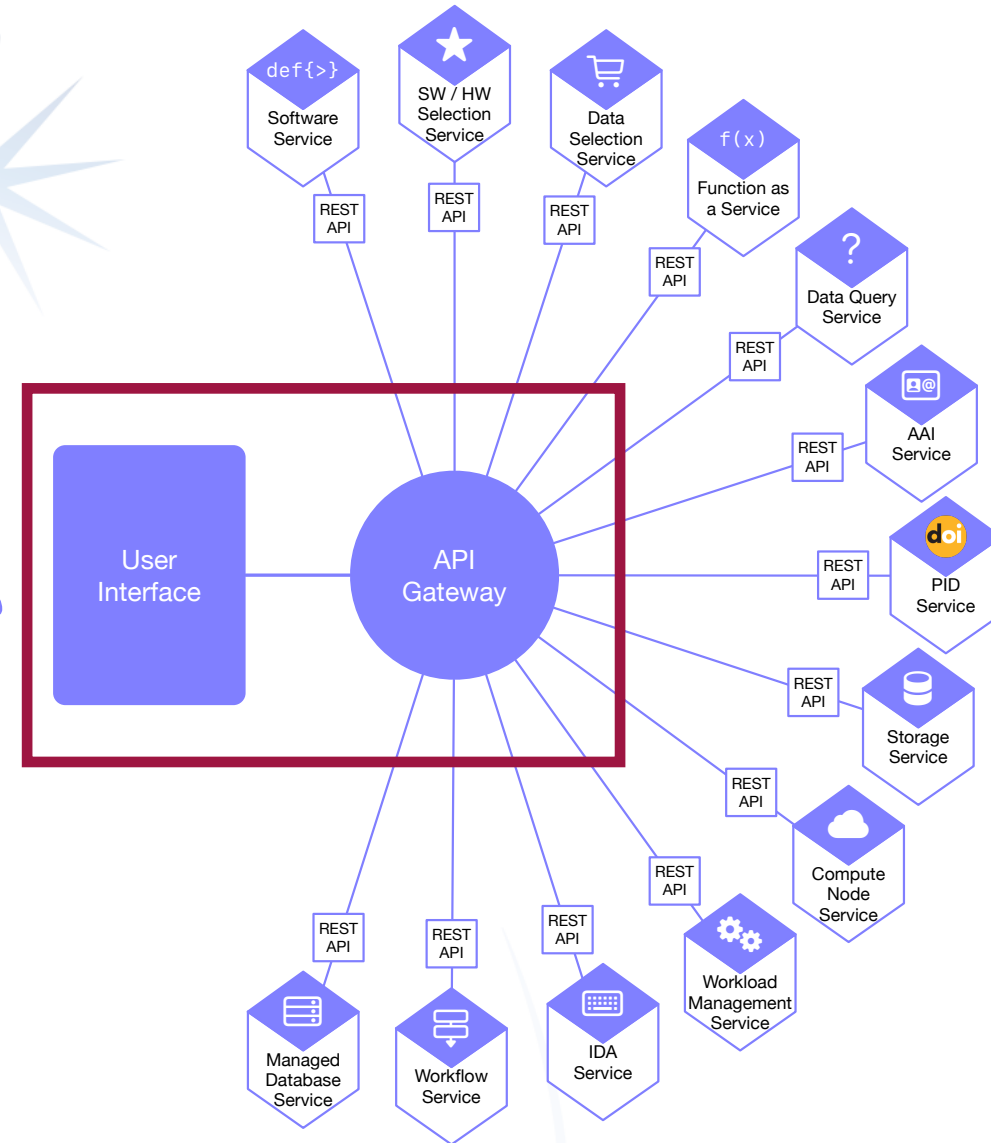
ESAP as a *toolkit*; integration with ESFRIs, EOSC

- WP5 is not resourced to operate or provide science platform services, except for development systems in support of

- The ESAP deliverables are oriented around enabling ESFRIs, other projects, and institutions to *deploy their own science platforms*.

- Individual ESAP instances can be e.g. registered with the EOSC Portal.

RI engagement

- Wide representation of RIs in WP5.

- ESAP is *customizable for different RIs.*

- The partners contribute to the development of a common core — the UI & API gateway — and collaborate to develop plugins to address their particular needs.

- Examples (*not exhaustive*):

  - JIVE, FAIR, SKA: Jupyter notebook integration

  - SKA, CERN: Rucio & data lake access

  - LSST-EU, CTA: Software discovery & OSSR integration

WP: RI:	WP1, WP6 & Manag.	WP2	WP3	WP4	WP5	
CTA	▬	▬	▬	▬	▬	ESFRI PROJECTS
EST	▬		▬	▬	▬	
KM3NET	▬	▬	▬		▬	
ELT and ESO	▬			▬	▬	ESFRI LANDMARKS
FAIR	▬	▬	▬		▬	
HL-LHC and CERN	▬	▬	▬		▬	
SKA	▬	▬	▬	▬	▬	
JIVE	▬		▬	▬	▬	ERIC
EGO			▬		▬	Others
LSST-Europe	▬				▬	

WP5 Structure

- **Task 5.1**: Data Aggregation & Staging

- **Task 5.2**: Software Deployment & Virtualization

- **Task 5.3**: Analysis interface, work flows, and reproducibility

- **Task 5.4**: Integration with HPC and HTC infrastructures

- **Task 5.5**: Work Package Management

Task 5.1
Data Aggregation and Staging

○ *Provide ESAP users with the capability to access and combine data from multiple collections and to stage that data for subsequent analysis.*

○ Major RP2 results include:

- Release of the *shopping basket* system: the central metaphor for data management through ESAP.

- The *Data-Lake-as-a-Service* provides integration between ESAP-based analysis notebook functionality and the **WP2**/DIOS "data lake", including data browsing, file transfer, & latency-hiding content delivery based on XCache.

- Prototypes of *persistent storage for notebook environments* based on NFS and Ceph.

- Virtual Observatory standards supported in ESAP including SAMP and TAP. Prototype VOSpace implementation underway. Close cooperation with **WP4**.

- Teams from JIV-ERIC, CTA, FAIR have worked on exposing their archives through ESAP.

- Integration of Zenodo (**WP3**) and Zooniverse (**WP6**) query capabilities.

Task 5.2
Software Deployment & Virtualization

- *Make the codes, tools, scripts, and other packages developed in support of the various ESFRIs and other RIs readily available to ESAP users.*

- Major RP2 results include:

  - Two ESAP instances are deployed and maintained for developer, testing, and integration purposes. A demo instance (with no expectations on service availability) is accessible at https://sdc-dev.astron.nl/esap-gui/.

  - Various ESCAPE partners have deployed JupyterHub and BinderHub instances for testing and development as well as science uses. These systems have been integrated with ESAP, (some) with ESCAPE IAM, and with institutional, project, and RI-level compute & storage facilities.

  - A range of analysis software has been packaged for deployment through ESAP and/or other ESCAPE infrastructures. Much of this work is closely **linked to WP3** as software is registered in the OSSR. Highlights include:

    - A Jupyter-CASA kernel in support of radio astronomy applications;

    - Dockerized FAIR analysis applications (R3BRoot, CmbRoot);

    - CONCORDIA: a containerized system to execute the CORSIKA air-shower simulation system using the DIRAC workflow engine;

    - Containers providing notebooks and scripts for astronomical analysis, including a reproducible analysis workflow of HCG-16 (a group of galaxies) and analysis of data from the COROT and PLATO satellites;

    - Containerized systems for EGO/VIRGO data analysis;

    - Containerized pipelines for LOFAR (radio astronomy) data analysis.

Task 5.3
Analysis interface, workflows, and reproducibility

- *Integrate the data access services provided in Task 5.1 with the software from Task 5.2 to provide users with a coordinated approach to data analysis in the ESAP context.*

- Major RP2 results include:

  - Integration between ESAP and the **WP3** OSSR. Users can search the OSSR from within ESAP, identify relevant software, and dispatch it to appropriate compute facilities for (interactive) execution. This service is currently based around BinderHub; additional execution capabilities are under development.

  - Integration between Jupyter notebooks and the ESAP shopping basket (Task 5.1). Notebook users have access to a convenient API for working with the basket contents. This work was carried out in conjunction with **WP6**.

  - Work is underway on a managed database service for use within ESAP. Users can choose to retrieve tabular data from archives directly into their own database, rather than adding to the Shopping Basket. They can then use the full power of SQL to perform complex analytics on their data. This is currently an early (but functional) prototype.

  - Also in prototype for is a system for managing, archiving, and issuing persistent identifiers for notebooks which are used for published analyses. This is currently being developed with the EVN archive, and has not yet been integrated with ESAP.

Task 5.4
Integration with HPC and HTC infrastructures

- *Make it possible for ESAP users to deploy their workflows and analysis jobs at scale on a range of HPC and HTC infrastructures.*

- Major RP2 results include:

  - Initial integration with the DIRAC system.

  - Experimentation with Apache Hive and Hadoop. This is currently used to support workflows at IFAE; full integration with ESAP may follow, pending successful outcomes and availability of resources.

  - Experimentation with batch processing for the PANDA experiment at GSI, including use of the **WP2** data lake.

  - Collaboration with **WP4** to develop a technical note exploring concepts for a web-service interface providing a simple way to discover and access computing services. This is an important step towards *common, interoperable standards* emerging from the ESCAPE project.

  - The *Rosetta* system, developed at INAF, is in the process of being integrated with ESAP as a way of managing compute jobs.

Task 5.5
Work Package Management

- *Coordinate the WP5 effort; manage interactions with other ESCAPE WPs; build and maintain connections with related work and projects in the community.*

- Working cadence:
    - "All hands" WP5 meeting every month to address major updates & stakeholder inputs.
    - Weekly technical meeting, focused on code core platform development.
    - Ad-hoc working groups formed to address key topics as needed.

- Links with other work packages through working groups and active discussion between key institutions (CERN/**WP2**, FAU/**WP3**, U. Edinburgh/**WP4,** OU/**WP6**).

- Presented ESAP & ESCAPE work at PUNCH4NFDI, ADASS, EAS2021, European Data Provider Forum, Low-latency Alerts & Data Analysis for Multi-Messenger Astrophysics, European Researchers' Night, ...

- Close links with Rubin Science Platform through CNRS-LAPP; multiple instances of RSP deployed for testing purposes.

ESFRI Alignment:
Summer 2021 workshop, MS31/D5.3

- The *Second WP5 Workshop* (2021-08-05) presented the current status of ESAP development & solicited input from the community.

  - http://indico.in2p3.fr/e/SecondWP5Workshop for agenda, materials, etc.

- D5.3 summarizes the results of that workshop, and compares the current status of ESAP to:

  - *Requirements*, captured in D5.2 (Detailed Project Plan);

  - *Use cases*, captured in the Project Platform.

- D5.3 made a number of suggestions for minor "course corrections", but concludes that the overall direction of travel and functionality is appropriate.

Deliverables

Number	Title	Due	Completed
D5.1	Preliminary report on D5.1 requirements for ESFRI WP5 science analysis use cases	M6 (July 2019)	M6 (July 2019)
D5.2	Detailed project plan for WP5	M9 (Oct. 2019)	M9 (Oct. 2019)
D5.3	Performance assessment of initial Science Platform prototype	M31 (Aug. 2021)	Submitted M31
D5.4	Final assessment of the performance of the Science Platform prototype and plan for deployment of production version within the EOSC	M42 (July 2022)	-

Milestones

Number	Title	Due	Completed
MS27	First WP5 workshop on Science Platform design and requirements	M4 (May 2019)	M4 (May 2019)
MS28	Review of preliminary report on requirements for ESFRI science analysis use cases	M5 (June 2019)	M5 (June 2019)
MS29	Initial science platform prototype with discovery and data staging	M22 (Nov 2020)	M22 (Nov 2020)
MS30	Deployment of initial set of ESFRI software on prototype platform	M23 (Dec. 2020)	M24 (Jan. 2021)
MS31	Second WP5 workshop to analyse prototype performance	M31 (Aug. 2021)	Submitted M31
MS32	Integration of Science Platform with OSSR repository	M32 (Sep. 2021)	Submitted M32
MS33	Integration of Science Platform with Data Lake expanded prototype	M24 (Nov. 2021)	Submitted M24
MS34	Delivery and integration of new ESFRI visualization and analysis tools	TBC	-
MS35	Final WP5 ESFRI user training workshop on the Science Platform	TBC	-

Risks Analysis: Grant Agreement

Risk	Description	Current Status
17	Poorly defined requirements on Science Platform functionality to support ESFRI science	• Primarily addressed through D5.1 & Project Platform, which extensively document the use cases being addressed. • Ongoing engagement of ESFRIs with WP5 to refine and address use cases. • Successful summer 2021 workshop validated project goals with ESFRIs.
18	Lack of commonality between ESFRI requirements	• Modular design of ESAP makes it possible to service multiple ESFRI needs from a common core.
19	Lack of availability or immaturity of underlying core EOSC services	• Development and test is isolated from EOSC services. • Active engagement with the EOSC Future TSPs will drive EOSC integration.
20	Insufficient ICT infrastructure available to support Science Platform deployment and testing	• Adequate infrastructure is available thanks to resource made available be a variety of project partners (thank you!).

Risks Analysis: Detailed Project Plan

- Lower-level risks are identified in D5.2 (Detailed Project Plan).

- "Retention of key staff, with domain knowledge" is (partially) realised through extended leave of WP5 Technical Lead.

  - Mitigations put in place included proactive engagement of the WP Coordinator, development of a common ESAP vision, and formation of a regular technical leadership team.

  - Milestone MS31 and Deliverable D5.3 were delayed relative to the original scheduled, but delivered in August.

- No other major concerns.

- We continue to monitor and proactively respond to risks.

Future plans

- *Focus on integration and stabilization of the various capabilities than have been developed across the work package and the ESCAPE ecosystem.*

- Upgrade to the underlying ESAP architecture to better support asynchronous services.

- Integration of Rosetta and other execution engines, supporting a wider range of compute and analysis task types.

- More advanced selection of software through WP3/OSSR, and support for intelligent dispatching of compute tasks.

- Deploy a batch processing management system based on DIRAC.

- Complete integration of the various data and analysis services which have been developed by project partners.

- (Continue to) engage with ESFRIs on ESAP deployment and service provisioning.

Conclusions

- ESAP provides a central coordinating element across the diverse ESCAPE ecosystem, bringing together work from WPs 2, 3, and 4 into a single user-facing system.

- During RP2, the WP5 team has made important technical progress...

  - ...on development of the ESAP system itself.

  - ...on packaging and providing a range of scientific analysis capabilities.

  - ...on engaging with the other work packages.

- Over the next several months, we will focus on consolidating and integrating this work so that we deliver a compelling package to all ESCAPE ESFRIs.