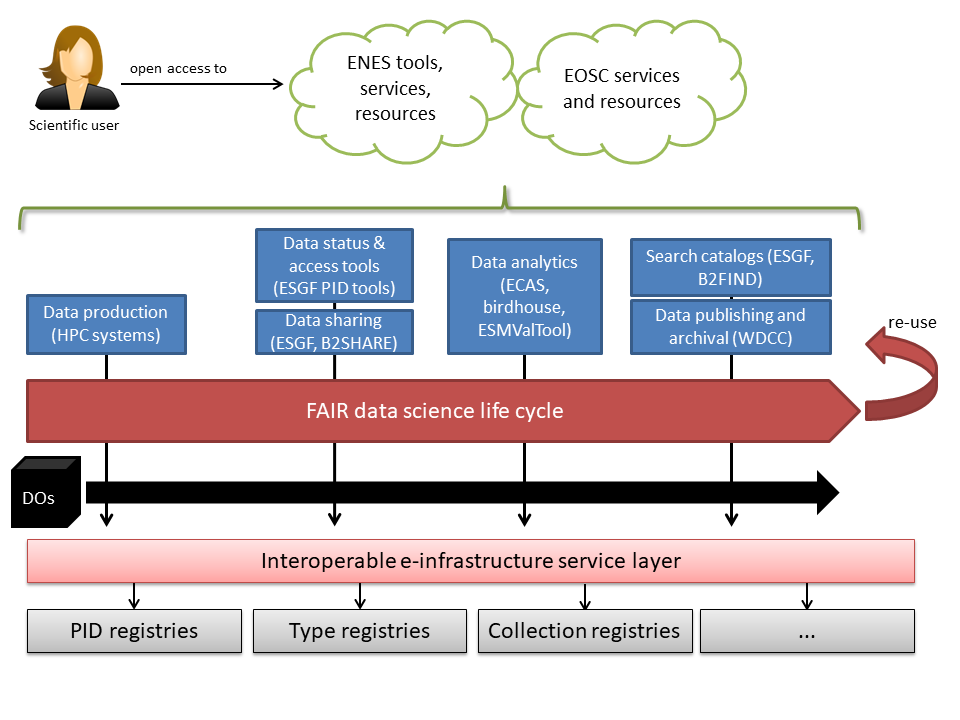
# The ENES case: Leveraging Kernel Information for process automation

*Tobias, …*

In Earth System sciences and its supporting infrastructures, we see an increase in the number, volume, complexity and diversity of digital research data objects. The realization of the FAIR vision depends on moving along e-infrastructures, including those relevant for ENES, towards providing the necessary functionalities for better automation of processes related to data, and establishing better and more automated ways to capture and manage provenance metadata.

Additional services also potentially increase the e-infrastructure workload. There are multiple options to deal with this, including hiring more staff, but also by making the processes more efficient. Therefore, a promising suggestion to realize the benefits of FAIR data is to increase the automation in the e-infrastructure tools and services that are part of the day-to-day business processes. Automation presents a strong opportunity, in particular, for reusability or at least verifiability of workflows, by capturing provenance information autonomously at multiple workflow steps and bringing this together to provide FAIR benefits through added-value services such as viewers, improved service chaining, expert systems and so on.



As shown in the diagram, users will interact with ENES and potentially also cross-disciplinary EOSC tools and services in order to work with resources, including data. Specific service examples, highlighted in blue, can be used by users at different stages of the data science workflow. Note that there may be different users or users in different roles interacting with a data object at different stages, and they may carry with them different knowledge.

In particular, the following examples highlight the role and interaction within the workflow:

* HPC systems will be used to run climate simulations, using models, based on experiment specifications. The resulting data are typically in netcdf format and massive in volume. They may be accompanied by comparatively extensive and well-standardized metadata in case of large community projects such as CMIP, but more often, metadata are scarce, not standardized, and potentially not quality controlled.
* Data will be shared with immediate peers or larger parts of the community for further analysis or review. The Earth System Grid Federation (ESGF) is the prime example used by the community. At this stage, support tools may be used to assess the status of data at hand (e.g., ESGF PID tools).
* Data may be analyzed using a variety of tools and services, and they shall be enhanced to capture provenance information. Examples include:
  + The ENES Climate Analytics Service (ECAS), developed as part of the EOSC-hub project, offering a JupyterHub environment with computing and data sharing capabilities
  + Birdhouse, a set of easy-to-use web processing services, offering simple analysis tasks on data
  + ESMValTool, an analysis toolbox used, e.g., in CMIP6
* Finally, data become mature enough to be archived, e.g. in WDCC, and be included in search catalogs such as the ESGF search or the more general B2FIND.

These services interact with digital objects (DOs) at different stages of the life cycle. The interaction with DOs should be done through an intermediate layer of e-infrastructure services, whose services and service components must be interoperable through specified interfaces, protocols, metadata standards etc. The interoperable service layer is able to satisfy demands by relying on a set of registries for PIDs, types and object collections, but also other bottom-layer service components.

Kernel Information provides the capability to bind the information together that is accumulated from the different services into the registries. The channeling through the interoperable service layer is essential to realize the automation benefits, e.g., for assembling knowledge graphs and perform service chaining between data sources, data transformers and VREs. The service layer presents a strong opportunity to gather provenance and combine provenance information from multiple workflow stages to be exploited by end-user facing services.

Key requirements for provenance-specific aspects include that provenance gathering should not impede the user’s default workflow, should be transparent to the user, but also take information from users that are active at different life cycle stages. For instance, a user who generates data on a HPC system may be entirely different from a user who uses a VRE for data analysis, and yet different again from a user who explores results or components to re-use in a catalog or from publications. Yet, only the specific user active at each stage has the necessary knowledge to provide the information in the first place.

Interaction with Kernel Information may therefore happen at all stages, by different services contributing different “snippets” of information to build up a full PID Kernel Information record. The primary storages of the metadata that is brought into Kernel Information records are diverse and include, e.g., the ESGF metadata catalog, databases in the analytics and HPC environments, and the search catalogs.

## Requirements for PID Kernel Information profile management

* Primary actions for a Kernel Information profile include: Create, Read, Update of a profile.
* When updating a profile, the old version should remain valid and persist. It may be marked as outdated, discouraging any new records from being created along it, but services should continue to be able to read it and use it for verification of Kernel Information records.
* Individual elements of Kernel Information records should be registered in a Kernel Information Element registry, e.g., a Data Type Registry instance, and they should also follow strict versioning policies.
* Each Kernel Information record should include a policy field (already specified in the RDA Recommendation) that tells a machine how an object may or may not change in the future. This information should be standardized across disciplines.
* ENES should be able to maintain community-specific profiles, but also relate to generic profile(s) that go beyond the community. ENES should have full authority to change its profiles, constrained by guidelines on, e.g., versioning.