Feedback from the WP5 Interactive Data Analysis WG

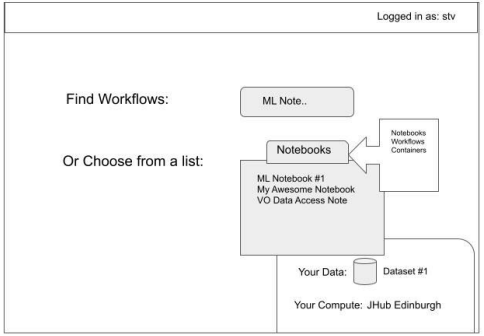
Stelios Voutsinas
ESCAPE WP5
University of Edinburgh

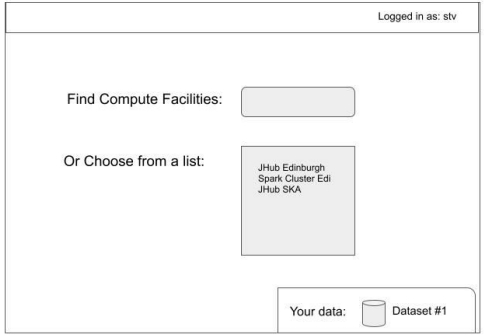
WP5 Working groups

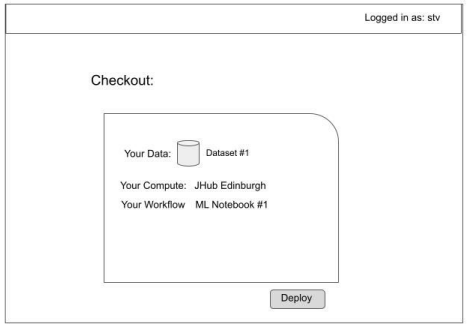
- Data discovery & Query
- **Interactive Data Analysis (IDA)**
- Data Staging (Rucio)
- Batch Processing

WP5 IDA Overview

- Interactive Data Analysis WG aims to develop the component of ESAP that will:
  - *Allow scientists to interactively discover, analyze & visualize data through the use of known & curated software & ESCAPE compute facilities*
- ESAP:IDA will be a Hub rather than a compute service
  - Discovering datasets, software (& workflows) and compute facilities, and allow users to connect all three via automated processes.
- Bring Code to the Data
- Bring Code & Data to the Compute Facility







ESAP IDA: UI Design Mockup

Current Status

- Initial plans in place, but effort not yet in full force on the IDA part of ESAP
- As a first prototype, focusing on JupyterHub & Notebooks
  - Notebook = Workflow
  - JupyterHub = Compute Facility
- Basic Demonstrator & Initial REST services developed
  - Hard-coded Service URLs & Workflows
  - Sample Github project for Notebook that deploys at MyBinder
- Integration with WP3 in progress for consuming the OSSR Software Repository

Upcoming Plans

- Complete Integration with OSSR for workflow & software discovery
- Develop Compute Facility discovery / registration
- Develop Auth flow between ESAP & Compute facilities.
- Enable automated staging of Workflow (Notebook) & Data at site (JupyterHub)
  - JHub REST API
  - Inject code for staging data if applicable
- Lots of Technical issues to solve..

WP5 & VO/WP4 Integration in ESAP

- Data Discovery & Query WG
  - Use of VO Registry for Service Discovery
  - TAP (& ObsTAP) Queries for Data Access
  - SAMP for Data Exchange between browser & local tools
- Interactive Data Analysis WG
  - Use of VO Tools for Data “Staging” (Planned)
  - VO Tools for Data Visualization
  - Example Notebooks with VO Discovery & Access as curated Workflows
- Rucio WG
  - VOSpace with Rucio implementation

Interoperability (IVOA) Topics/Issues

- **Use of SAMP with HTTPS** (Data Discovery & Query)
- **Standardizing Software Description & Metadata** (Interactive Data Analysis)
    - WP3 are planning on using Codemeta for defining the software metadata & Zenodo to be the Software Repository.
        - Codemeta: https://codemeta.github.io
        - Zenodo: https://zenodo.org
- **Standardizing access and discovery of Software & Workflows (Interactive Data Analysis)**
    - How does ESAP discover and offer curated software and/or workflows to users?
        - In the current plan this means ESAP harvesting metadata Zenodo REST API.
    - What metadata is required to describe all different types of Software & Workflows

Interoperability (IVOA) Topics/Issues

- **Curation of Notebooks (Interactive Data Analysis)**
  - Notebooks as a form of software?
  - Curate them allowing other to discover and reuse them with 100% repeatability.
  - What metadata do we need to describe a notebook to do this?
    - (i.e. version, author, required libraries, required hardware specifications such as storage required..).

Interoperability (IVOA) Topics/Issues

- **Standard Discovery & Access to Compute Facilities (Interactive Data Analysis)**
  - Requires a standard way of interacting with these services, both for discovery but also for automation of tasks. (i.e. automatically deploy a workflow and stage data at a JupyterHub Facility).
  - JupyterHub -> REST API
  - Possible to have more abstract metadata on a compute facility? (Notebook service, VM access, Spark cluster, etc..),
  - For example
    - What type of authentication is required?
    - What software libraries are available?
    - How many resources a user can have?
    - Access protocols (ssh, http..)?
    - Whether a known user has access and others?

Codemeta

- Create minimal metadata schemas for science software and code.
- Standardize the exchange of software metadata across repositories and organizations.

```
{
    "@context": "http://schema.org",
    "@type": "Code",
    "author": [
        {
            "@id": "http://orcid.org/0000-0002-3957-2474",
            "@type": "Person",
            "email": "arfon@github.com",
            "name": "Arfon Smith"
        },
        {
            "@id": "http://orcid.org/0000-0002-7217-4494",
            "@type": "Person",
            "email": "kaitlin@mozillafoundation.org",
            "name": "Kaitlin Thaney"
        }
    ],
    "citation":
"http://dx.doi.org/10.6084/m9.figshare.828487",
    "codeRepository": "https://github.com/arfon/fidgit",
    "dateCreated": "2013-10-19",
    "description": "An ungodly union of GitHub and Figshare
http://fidgit.arfon.org",
    "keywords": "publishing, DOI, credit for code",
    "license": "http://opensource.org/licenses/MIT",
    "name": "Fidgit"
}
```

Zenodo

- General-purpose open-access repository developed under the European OpenAIRE program and operated by CERN
- Entries can be datasets, source code, publication etc..
- DOIs to identify each entry
- DOIs don’t change, but we can update versions
- Allows the creation of communities to gather together publications and entries under a similar subject/interest
- Allows harvesting in Machine-readable formats
    - Open Archives Initiative Protocol for Metadata Harvesting – OAI-PMH
    - oai_datacite and others

Summary & Open Questions

- WP5 IDA: Early in Development stages but clear need for:
  - Interoperable discovery & access to heterogeneous compute resources
  - Discovery & Access of Software & Workflows
- Notebooks treated as curated software?
- What can the IVOA take as lessons/ideas from the ESCAPE project?
  - Software description & Discovery?
  - Is there / Can there be a standard way for describing what a Compute Facility can do/run? Are there other use cases that would need this outside of ESCAPE?
- What else can ESCAPE WP5 re-use from existing VO standards & solutions?