WP3 OSSR – ESCAPE Repository

Thomas Vuillaume

ESCAPE WOSSL, 24/07/2020

ESCAPE Open-Source scientific Software and Service Repository

Goals

- Maximise software re-use

- Foster co-developments

- Identify open standards for software releases

- Investigate data mining tools and new analysis techniques

- Expose community-specific data services under the FAIR principles

→ Open-access repository to share scientific software, digital libraries do data analysis, datasets, documentation, tutorials and training activities to be included in the EOSC catalogue

Important note

Software **Repos**itory != **Develop**ment Platform

<u>Why is that important</u>

- To not re-invent existing tools, analysis
- **Combine** efforts
- Go further in the **multi-messenger** analysis domain
- Promote and ensure **open**, **reproducible science**, encouraging the implementation of the FAIR principles

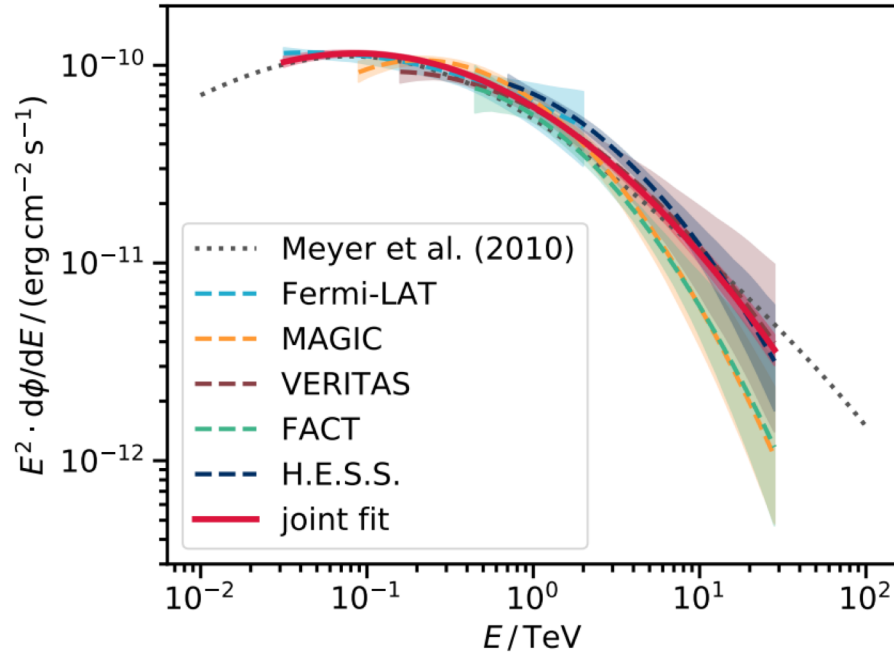# An example of open science project : The Crab bundle

Fig. 2. Crab nebula SED for individual instrument fits and from the joint fit. Single-instrument results are represented with dashed lines, the fit of all the datasets together, labelled as joint, is represented as a thick, solid red line. The shaded areas represent the SED error bands whose calculation is explained in Sect. 3.2. The dotted line shows the model in Meyer et al. (2010).
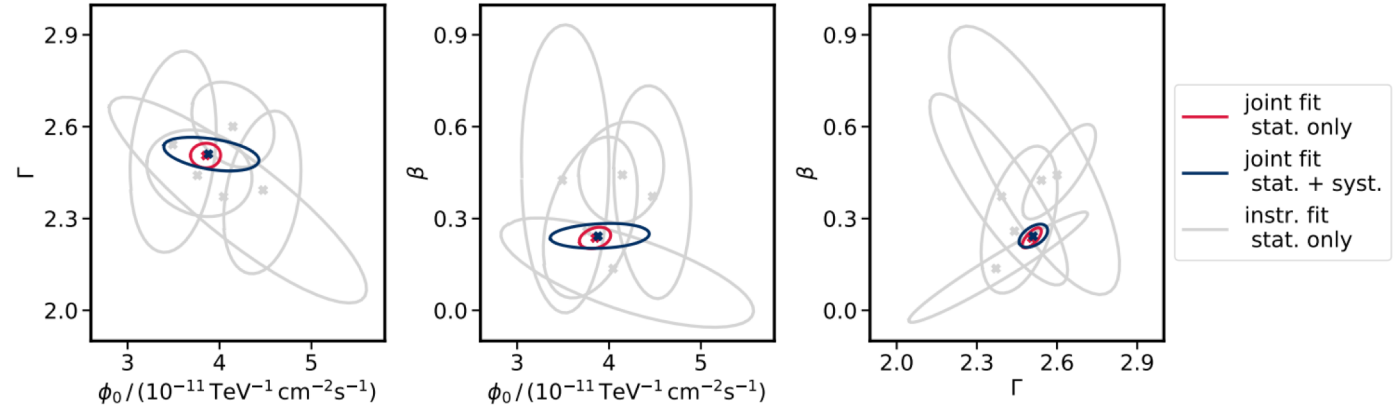

Fig. 5. Likelihood contours corresponding to 68% probability content for the fitted spectral parameters (ϕ_0, Γ, β), for the likelihood in Eq. (1) (red) and the likelihood in Eq. (5) (blue). Results from the individual instruments with the likelihood in Eq. (1) are shown in grey.

Multi-instrument gamma-ray analysis of the Crab Nebulae with Fermi-LAT, MAGIC, VERITAS, FACT and H.E.S.S.

Towards open and reproducible multi-instrument analysis in gamma-ray astronomy, C. Nigro et al, in A&A 625 (2019)

An example of open science project : The Crab bundle

The Crab multi-instrument gamma-ray analysis with MAGIC, VERITAS, FACT and H.E.S.S.
https://github.com/open-gamma-ray-astro/joint-crab/tree/v0.1
https://zenodo.org/record/2381863#.XkxcD5NKhhA

license

Cited by

Source code and data

Link to project and article

Cite as

An example of open science project : The Crab bundle

The Crab multi-instrument gamma-ray analysis with MAGIC, VERITAS, FACT and H.E.S.S.
https://github.com/open-gamma-ray-astro/joint-crab/tree/v0.1
https://zenodo.org/record/2381863#.XkxcD5NKhhA

- Binder to test and run analysis interactively online

- Docker to ensure reproducibility

ESCAPE Virtual Environment

Now imagine this as a standard,

in an integrated environment - with a single login,

allowing you to (re-)run (part of) any analysis,

with another dataset,

and easily publish your new results,

automatically giving credit to original analysis, datasets, workflows...

How ?

ESCAPE Virtual Environment

1. Promote and implement **FAIR*** principles

- **Findable** → Rich metadata, **unique** and **persistent** identifier

- **Accessible** → Metadata and data are **understandable** to **humans** and **machines**. Data is deposited in a **trusted repository**.

- **Interoperable** → Metadata use a formal, accessible, shared, and broadly applicable language for knowledge representation.

- **Reusable** → Data and collections have a **clear** usage **licenses** and provide accurate information on **provenance**.

2. Provide the infrastructure and services to foster FAIR (effortless) contributions

→ ESCAPE Virtual Environment

ESCAPE Virtual Environment
(a biased view)

ESCAPE Virtual Environment
(a biased view)

Jane as a new science idea

ESCAPE Virtual Environment
(a biased view)

Exchange documents — Collaborative Platform — Chat / Forum

Needs from developpers
- Answer questions and provide support to users
- Discuss with other developers
- Publish software / datasets — **Visilibility** / **Citable**

Continuous Integration and Development
Issues
Development Platform
Manage project

Comment / Ask questions
Rate
Re-use
Cite

Publish Releases

Long-term archive
Transparent use of the data
Findable data
Data Lake

ESCAPE Repository

Needs from users

- As been used / cited by
- Find associated publication
- Find associated datasets
- Find documentation / tutorials
- Research and filter

Find data
Build Container
IVOA

Publish analysis

query data

Run existing analysis
Modify existing analysis / workflow
Re-use analysis in other workflow

Re-run existing analysis with other dataset
Access and install software from repository
Analysis Platform

Interactive Test

Common A&A

Build new analysis

ESCAPE Virtual Environment
(a biased view)

ESCAPE Virtual Environment
(a biased view)

ESCAPE Virtual Environment
(a biased view)

Exchange documents — Collaborative Platform — Chat, Forum

Answer questions and provide support to users
Discuss with other developpers
Needs from developpers
Publish software / datasets — Visibility, Citable

Continuous Integration and Development
Issues
Development Platform
Manage project

Comment / Ask questions
Rate
Re-use
Cite

Publish Releases

Long-term archive
Transparent use of the data
Data Lake
Findable data

ESCAPE Repository

As been used / cited by
Find associated publication
Find associated datasets
Find documentation / tutorials
Research and filter

Find data

Build Container

IVOA

Needs from users

Discovery of software / services

query data

Publish analysis

Run existing analysis
Modify existing analysis / workflow
Re-use analysis in other workflow

Re-run existing analysis with other dataset

Access and install software from repository

Analy...

Interactive Test

Build new analysis

Common A&A

ESCAPE Virtual Environment
(a biased view)

Exchange documents — Collaborative Platform — Chat, Forum

Continuous Integration and Development
Issues
Development Platform
Manage project

Needs from developpers — Answer questions and provide support to users, Discuss with other developers, Publish software / datasets — Visilibity, Citable

Comment / Ask questions
Rate
Re-use
Cite

Publish Releases

Long-term archive
Transparent use of the data
Findable data
Data Lake

Repository

Build Container

IVOA

Find data

query data

Re-run existing analysis with other dataset

Access and install software from repository

Publish analysis

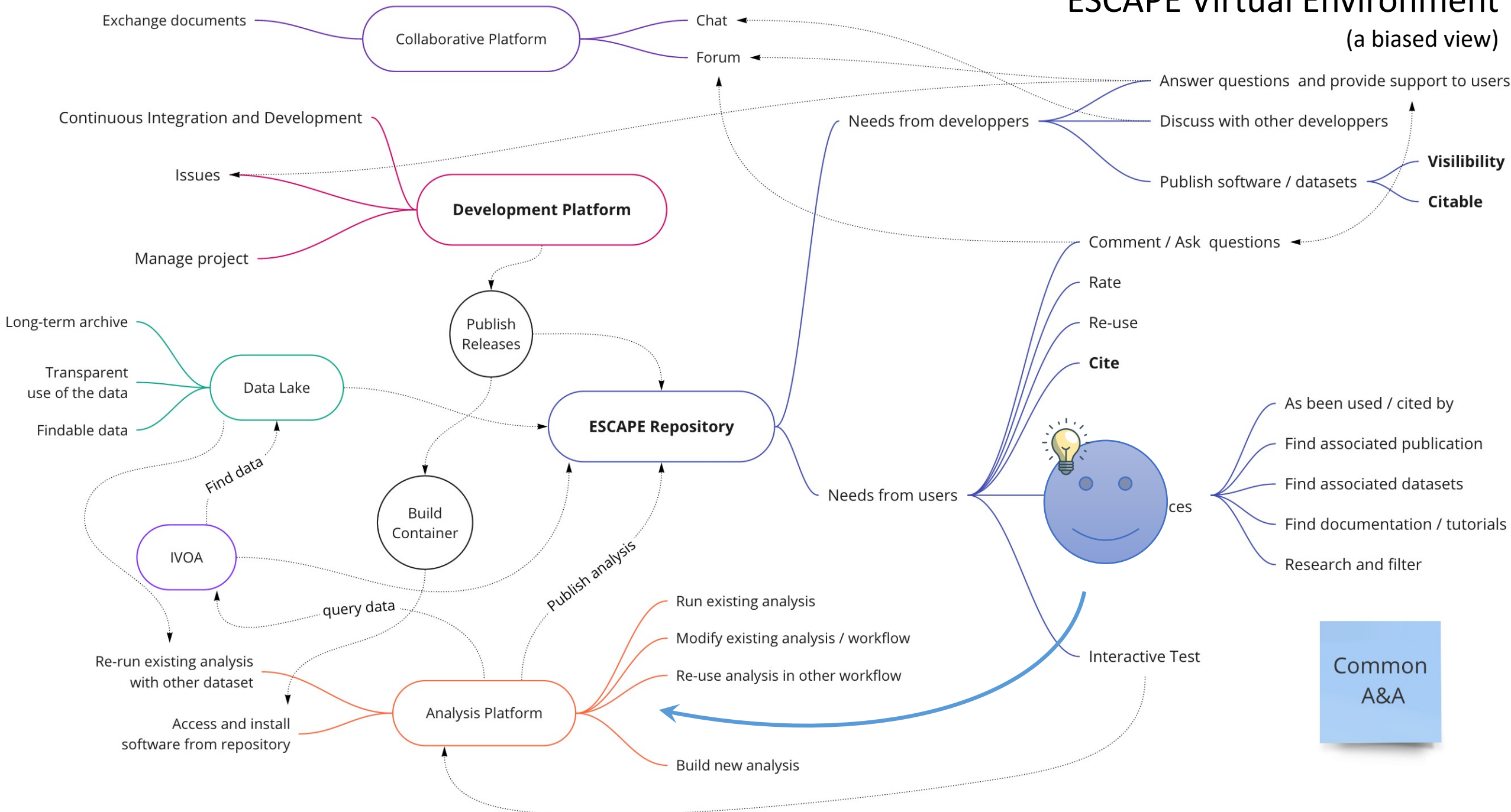Needs from users — Discovery of software / services — As been used / cited by, Find associated publication, Find associated datasets, Find documentation / tutorials, Research and filter
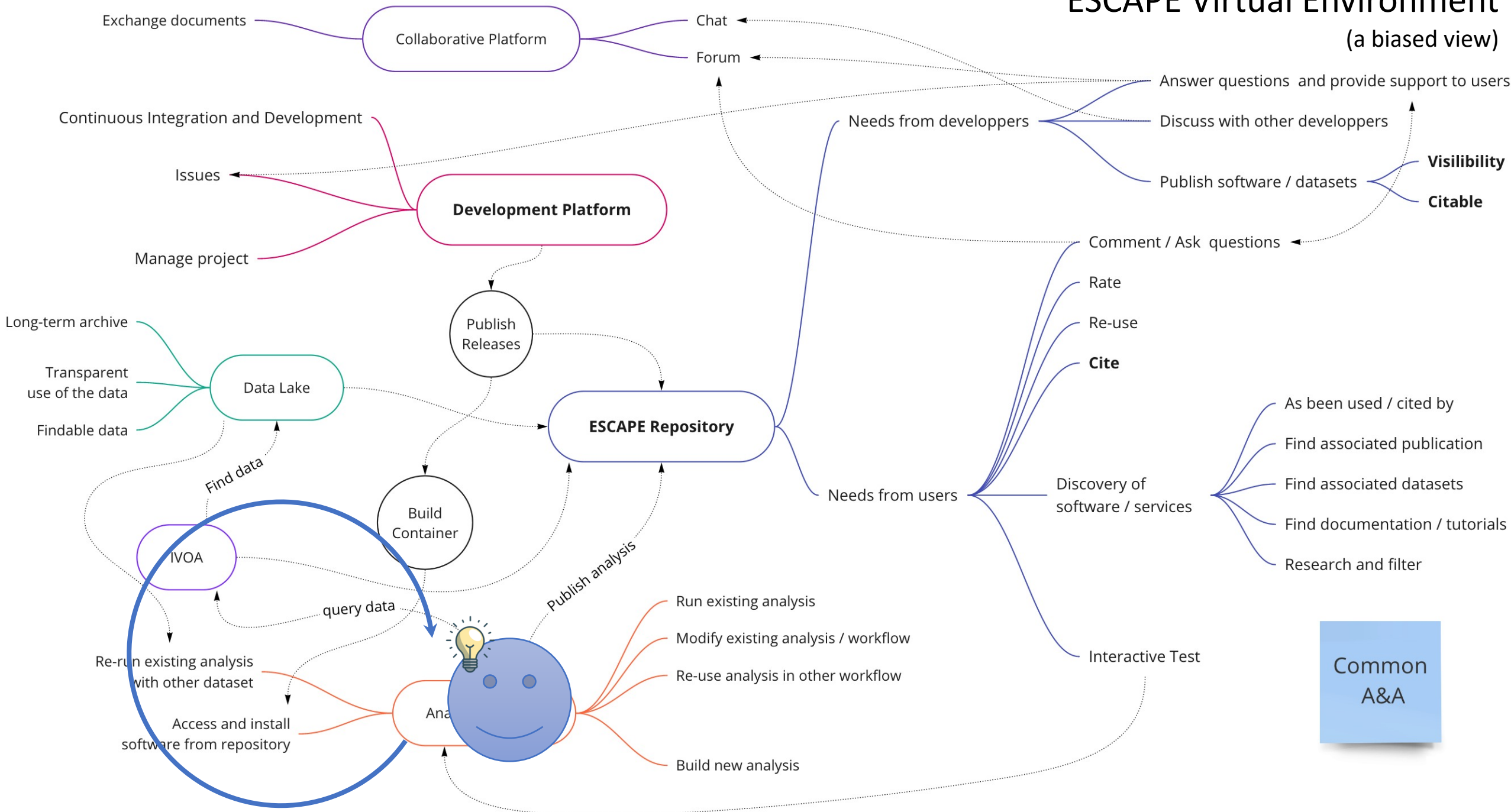
Interactive Test

Run existing analysis
Modify existing analysis / workflow
Re-use analysis in other workflow
Build new analysis

Analy...

Common A&A

Jane's new idea could be quickly tested in a virtual environment,
with the required software pre-installed,
and with access to previous and new datasets.

In the process, she not only made new discoveries,
she became an active developer of the tools,
and published openly her analysis,
allowing other to cross-check, re-use, cite and build on top of it.

→ the repository must be **more than an archive**

Technical implementation ?

Partners survey on the repository

- *Do you have one or several repositories in place within your ESF/RI and how are they implemented?*

- *What features do you expect from the OSSR repository?*

- *How would your ESF/RI transfer the software and services to the OSSR (transferring the full development, adding stable releases or similar?)*

- *What software licenses, versioning and means of preservation do you employ?*

- *How does your ESF/RI implement the FAIR principles to software, services and relevant data to test those, what means are in place for long-term preservation, maintenance and curation?*

- *What features do you expect from the OSSR repository?*
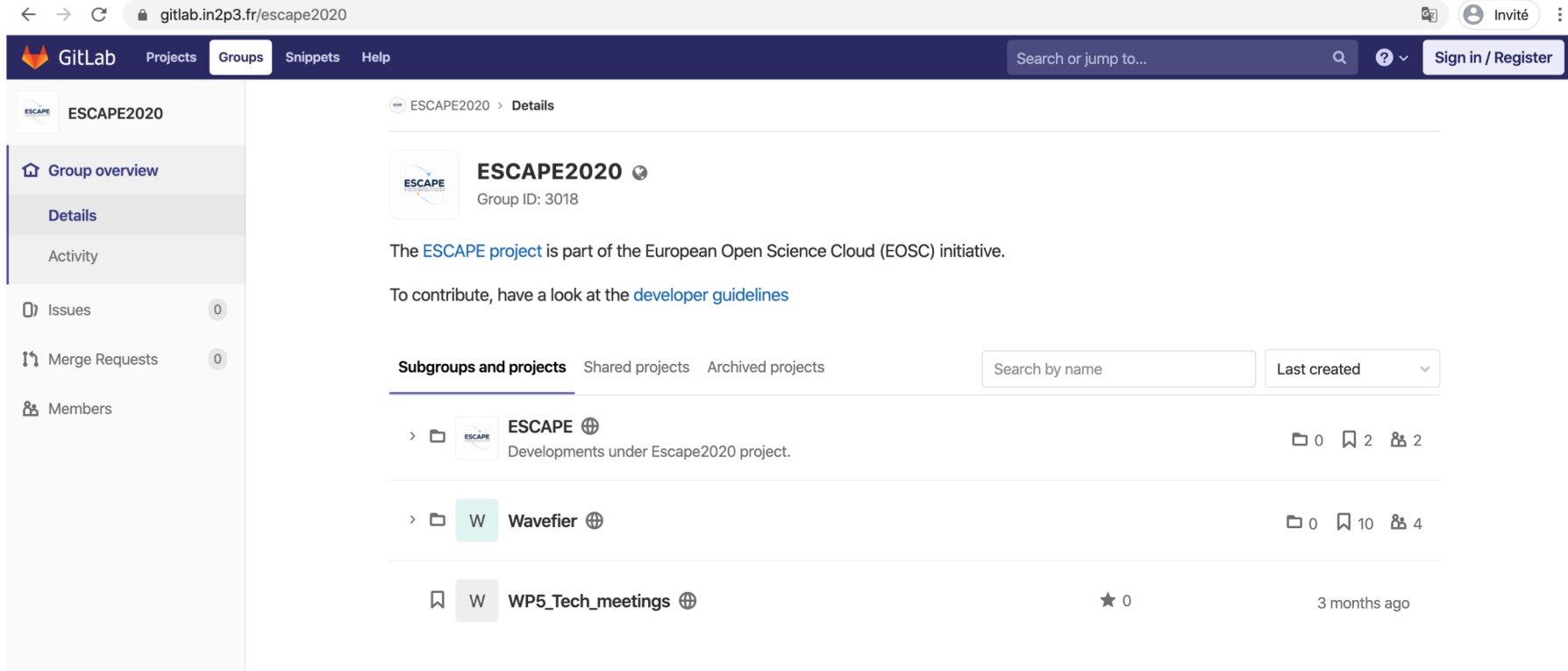
- *…*

Development platform: Gitlab
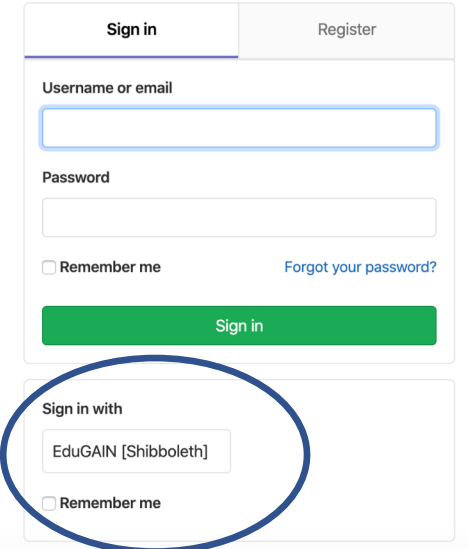
https://gitlab.in2p3.fr/escape2020

Sign in with EduGain

Development platform: Gitlab

	+	-
Github or gitlab as a service	- Free for open source - No setup - Integration with lot of services - Huge community	- Data stored outside the E.U. - No control over the data & conditions of use - No control of accounts
Gitlab as a software	- Private projects - Control - Open source	- Setup, install, maintenance of the service - Costs - Less integration with other services

Development platform: Gitlab

	+	-
Github or gitlab as a service	- Free for open source - No setup - Integration with lot of services - Huge community	- Data stored outside the E.U. - No control over the data & conditions of use - No control of accounts
Gitlab as a software	- Private projects - Control - Open source	- Setup, install, maintenance of the service - Costs - Less integration with other services

Repository: concept implementation with Zenodo service

https://zenodo.org/communities/escape2020

ESCAPE 2020

keywords allow to filter by science project or research infrastructure

Zenodo as a service vs Zenodo as an application

	+	-
Zenodo as a service	- Free - No setup - CERN support - Secure, **very** long time, archive - Always up to date	- Size limitation to 50GB per upload (an agreement is possible for bigger datasets) - No multiple ownership as of today - No multiple curators as of today - Limited customization
Zenodo as a software	- Better branding - Own the data - Custom front-end - Custom A&A	- Need time and money to dev/setup/install - Need computing infrastructure - Need maintenance - Future ?

Zenodo as a service vs Zenodo as an application

	+	-
Zenodo as a service	- Free - No setup - CERN support - Secure, **very** long time, archive - Always up to date	- Size limitation to 50GB per upload (an agreement is possible for bigger datasets) - No multiple ownership as of today - No multiple curators as of today - Limited customization
Zenodo as a software	- Better branding - Own the data - Custom front-end - Custom A&A	- Need time and money to dev/setup/install - Need computing infrastructure - Need maintenance - Future ?

Note: Some of the limitations from « Zenodo as a service » will be overcome in the future or can be discussed with the Zenodo team through an agreement if needed

Under development or discussion

1. **Connect** the services to create an integrated environment

2. Define **metadata**
   - For humans
   - For machines

3. Define a **template** to upload software (<-> metadata)

4. Propose **license** and **provenance** model (<-> metadata)

5. **Feed** the repository with ESCAPE developments

6. Feed to repository with test **science projects** (see e.g. Dark Matter talk)

7. Provide a single entry point for science projects and ESCAPE virtual environment

See Enrique Garcia's presentations (monday) for technical details

Community Driven Repository, Organised by Science Projects

*Jane as a very big idea that needs a completely new development environment

Objective: ease co-developments and new cross-experiments science projects

- Starting a new Science projet (through validation) sets up a complete virtual environment

- Researchers / Institutes contribute to the science project by publishing software / workflow / data

- The contributions are validated by science project curators

- Users can search the repository or explore it through the science projects

Final word

- Aiming for an integrated virtual environment

- The repository hosts the software

- Central part of the environment

- Only as good as the software it is hosting
  - we need to provide very clear instructions to contribute
  - a lot of points yet to clear and discuss (metadata!)

See Enrique Garcia's presentations on Monday
for a status, more technical details, and a demonstration of the repository.

Questions?