UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'astronomie

EPFL

Astronomical **Open Research Data** Analysis Services

Volodymyr Savchenko

Digital Innovators Webinar
10 May 2023

Overview

- Multi-Messenger and Time-domain **Astronomy**: challenges and successes
- **Ecosystems for supporting science lifecycle**: data, process/workflow/service, and knowledge **stewardship**
- **Development environment for crowd-sourcing scientific workflows**
- **Knowledge Graphs**, Linked Data, Ontologies: **why** do we need them and **how** to use them
- Integration with **Publications**
- **Synergies** with other related developments within astronomy and other domains
- **Future** plans, hopes, open questions

Multi-Messenger Time Domain Astronomy

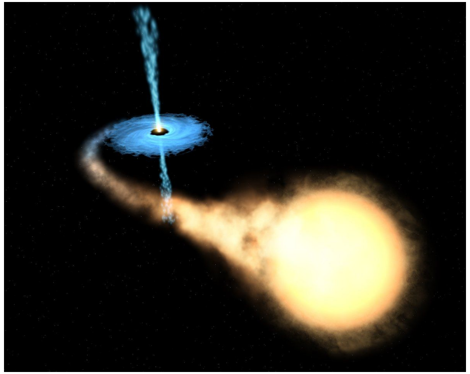
Exploding field!

Last decade key **new observables** were discovered, and conventional telescopes dramatically upgraded to match.

Number of alerts and **volume of data** we deal with **increased by couple orders of magnitude in the last years**, and several nearly-ready telescopes promise another comparable increase

"Just" a star

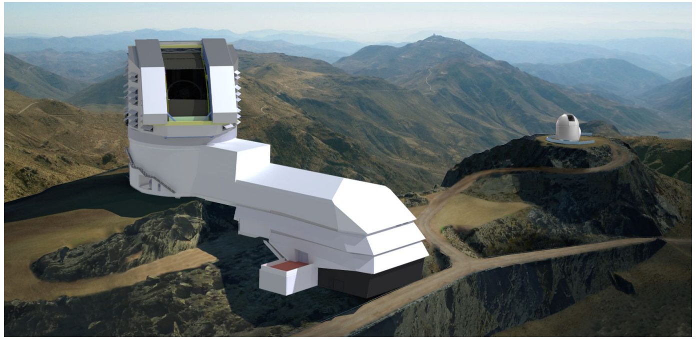
Star and black hole



Two neutron stars

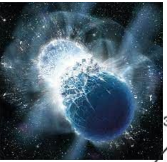
Radio

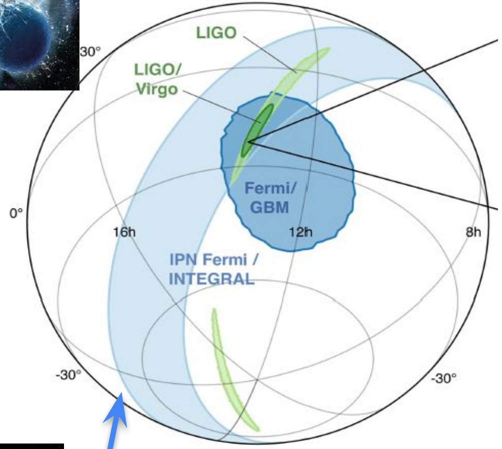
Visible

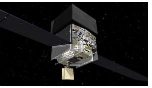


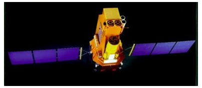
Gravity

Multi-messenger astronomy is collaborative and fast



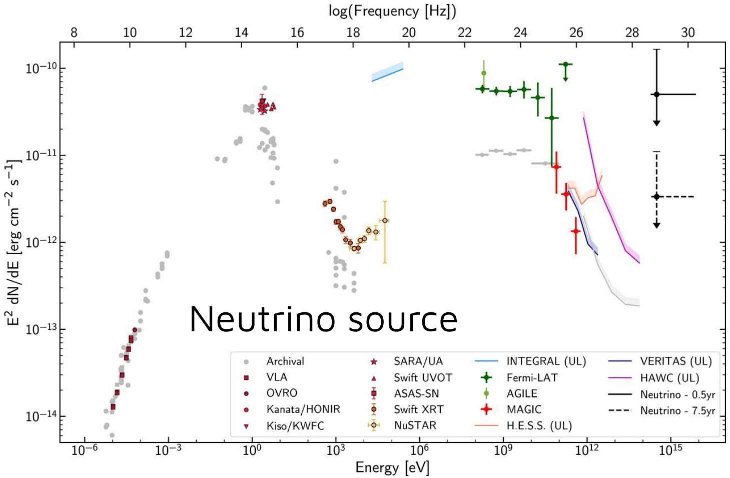


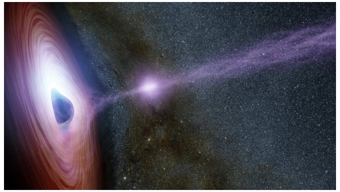




Fermi + INTEGRAL
"Triangulation"
unique multi-mission approach

VS+ 2017, LVC 2017





Our focus on **broad synergies** allowed us to take a leading role or contribute in some of the key recent discoveries in our domain:

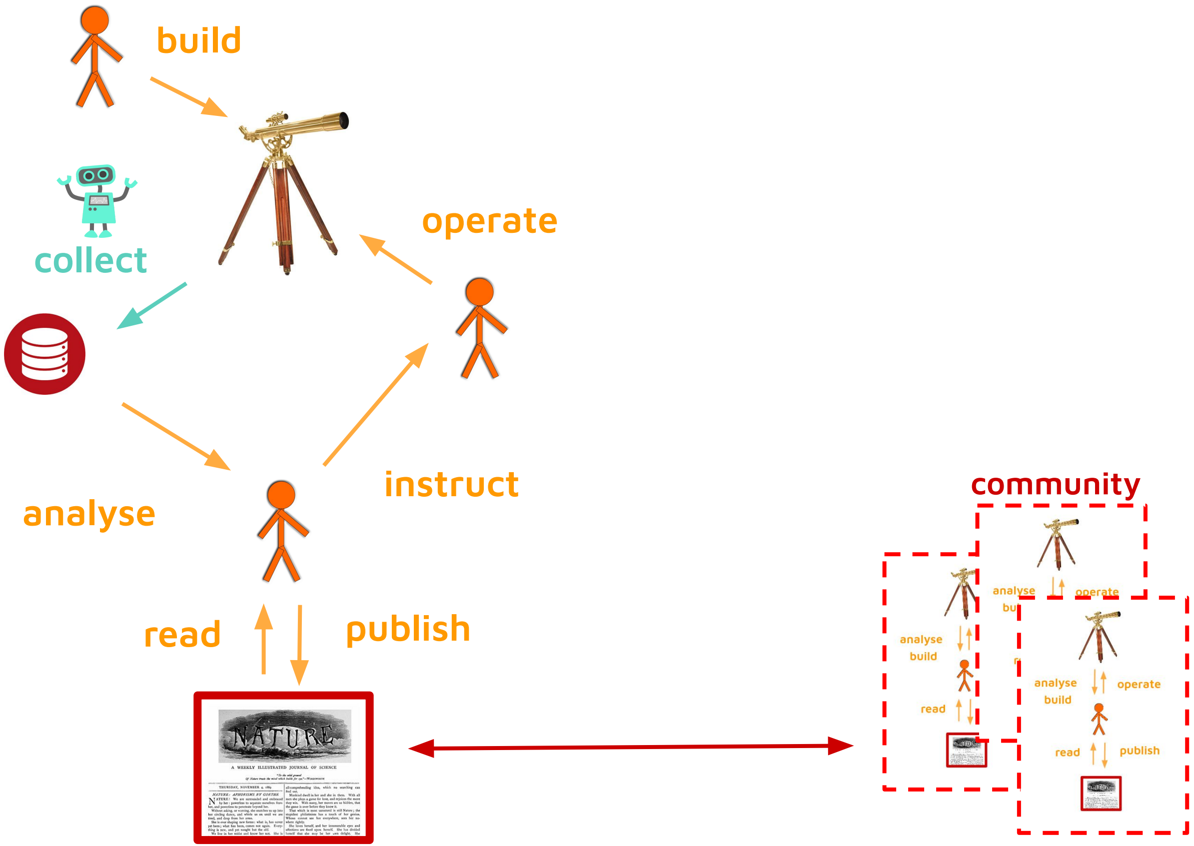
- Detection of the **first Gravitational Wave - Light Burst** coincidence (2017)
- First detection of light emission from **high-energy neutrino source** (2018)
- Discovery of the origin of a **Fast Radio Burst** (2020)

And it's getting worse: much more data, more diverse.

It's getting really hard to keep up with volume and veracity

Mostly-human Astronomy

- Reaction to sky: **slow**
- Reaction to papers: **slow**
- Trials (p-hacking): **uncontrolled**
- Publishing: **slow**
- Scalability: **bad**
- Creativity: **high**
- Communication: **nuanced but imprecise and slow**

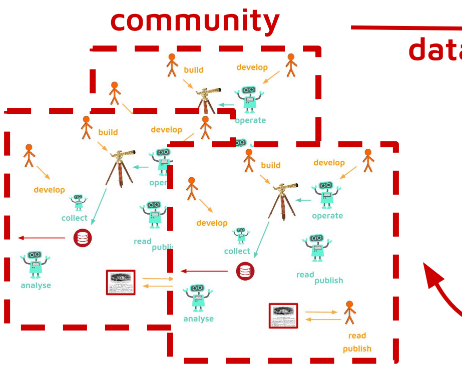


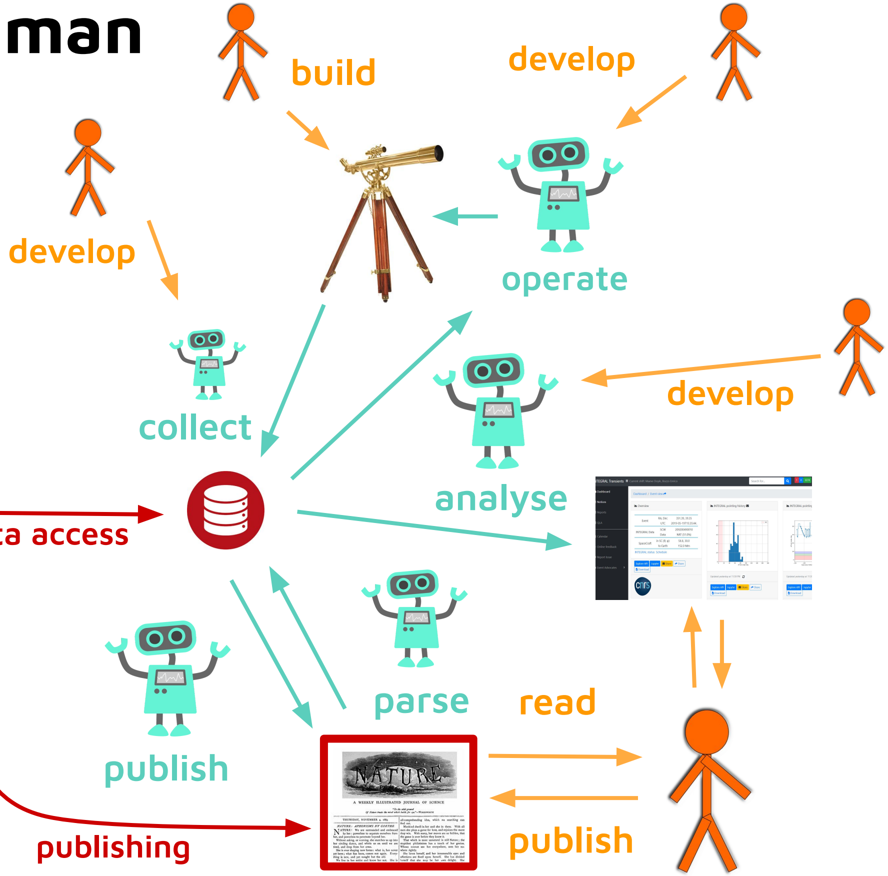
Human reaction and processing **is slow**, even if it's within even one person. But people are **smart**

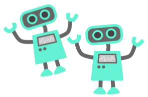
More robot, but even more human

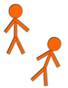
"Ideal" picture: most reality halfway there

- Reaction to sky: **fast**
- Reaction to papers: **fast**
- Trials (p-hacking): **controlled**
- Publishing: **fast**
- Scaling: **good**
- Creativity: **low**
- Communication: **precise, limited**









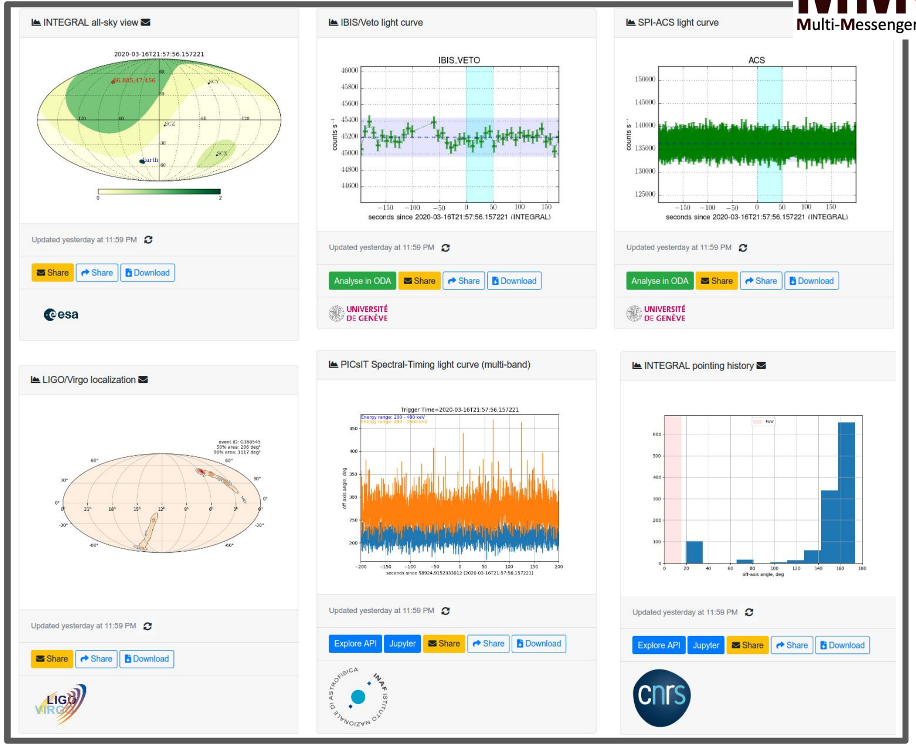
- **Making smart robots is hard**: always lacking **developers who are also research scientists.**
- If all is automated, **scientists have hard time seeing what's going on**, since **they do not speak robot**
- Robots are **fast**, but **lack creative reaction** in **new situations**.

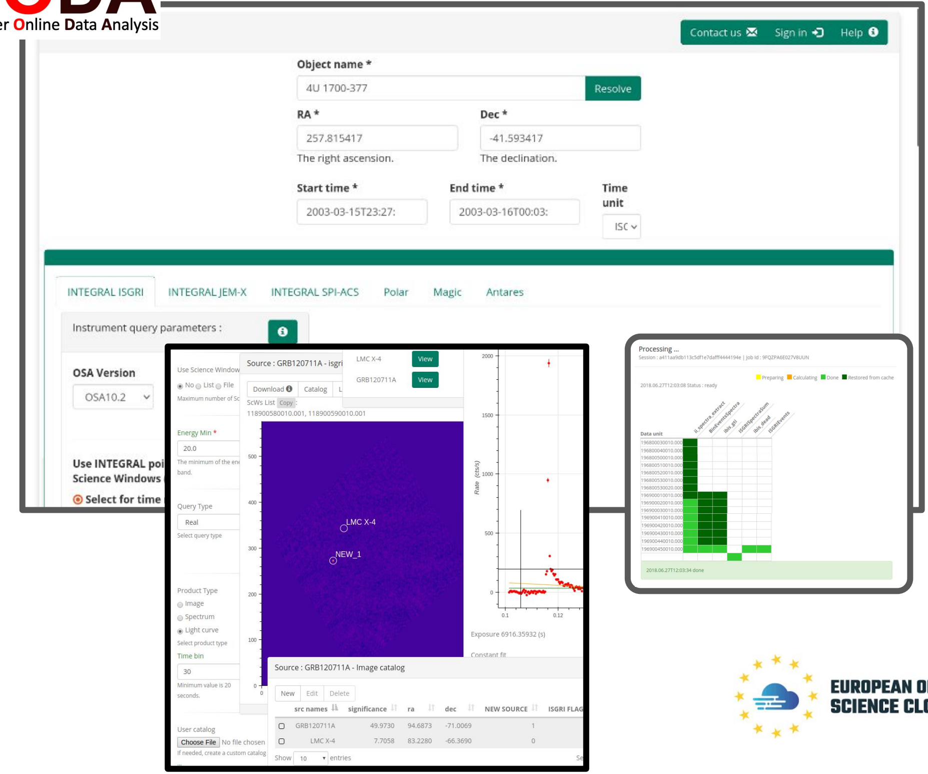
Tools for exploring, transforming research data

https://www.astro.unige.ch/mmoda

MMODA
Multi-Messenger Online Data Analysis

Deploy your own copy! https://github.com/oda-hub



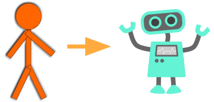


EUROPEAN OPEN SCIENCE CLOUD

Very hard to build these tools, need expert astronomers with state-of-the-art tool-building skills self.

https://marketplace.eosc-portal.eu/services/astronomical-online-data-analysis-astrooda

Development space: help scientists make robots



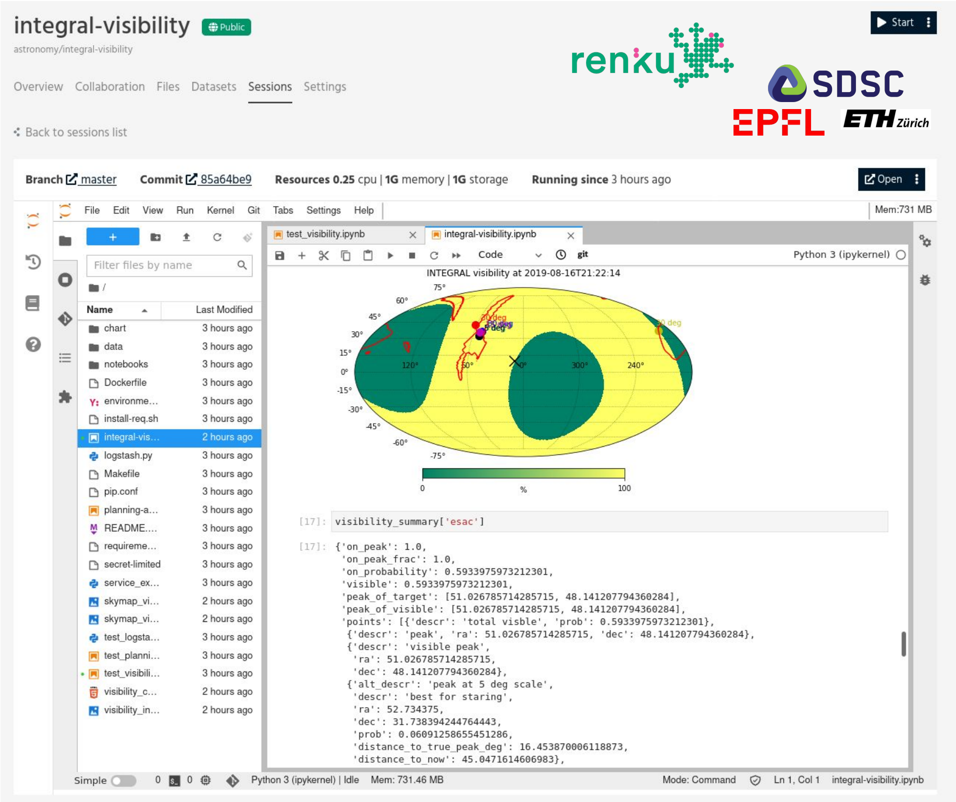
There are much **more scientists who can make a jupyter notebook than write organized code**.

JupyterHub(s), Google-collab, ESA DataLabs, Renku

- **Continuous integration and testing**
- Supports in **publishing of data and code** (e.g. in zenodo)
- Support in **annotation** for scientists and robots reuse with ontology terms

Develops and integrates metadata in a Knowledge Graph (see later)

This process creates a collection of notebooks and other workflows, but they are only really accessible interactively one-by-one

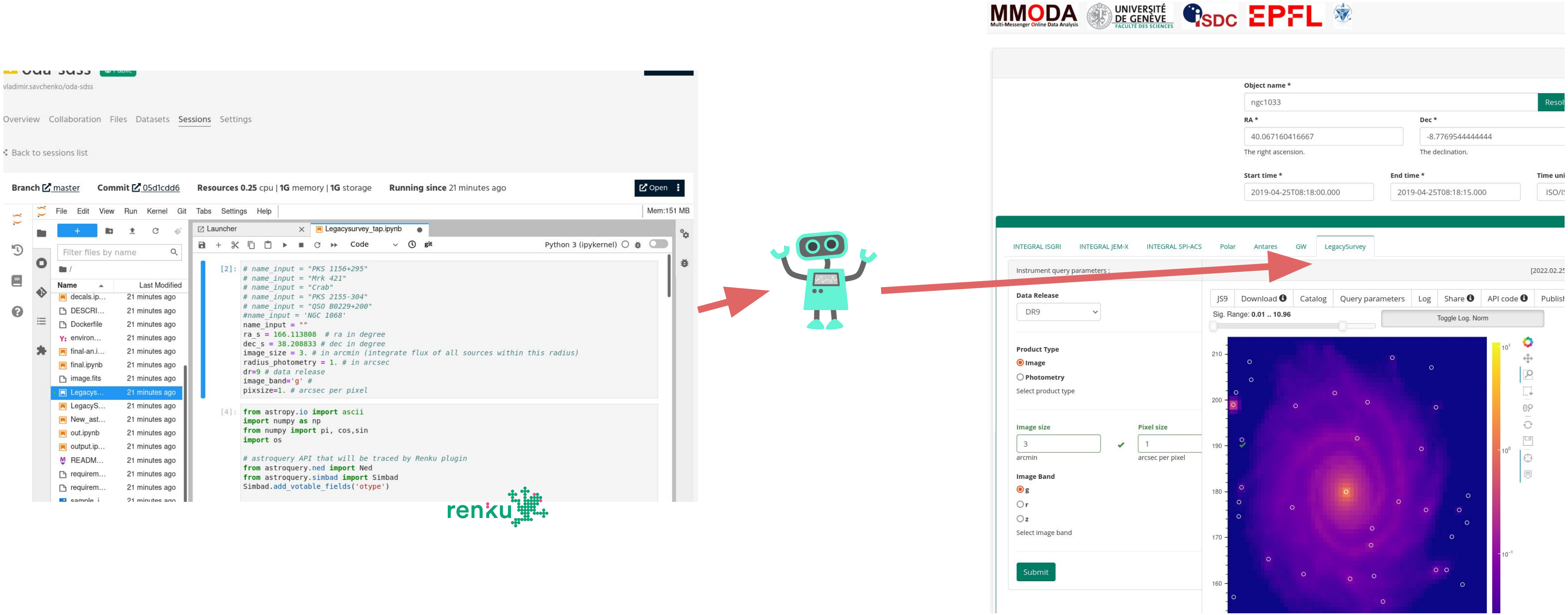


Making the developed workflow available as a web tool

jupyter may be easy, but sometimes we want just put parameters and click one button **in web interface.**

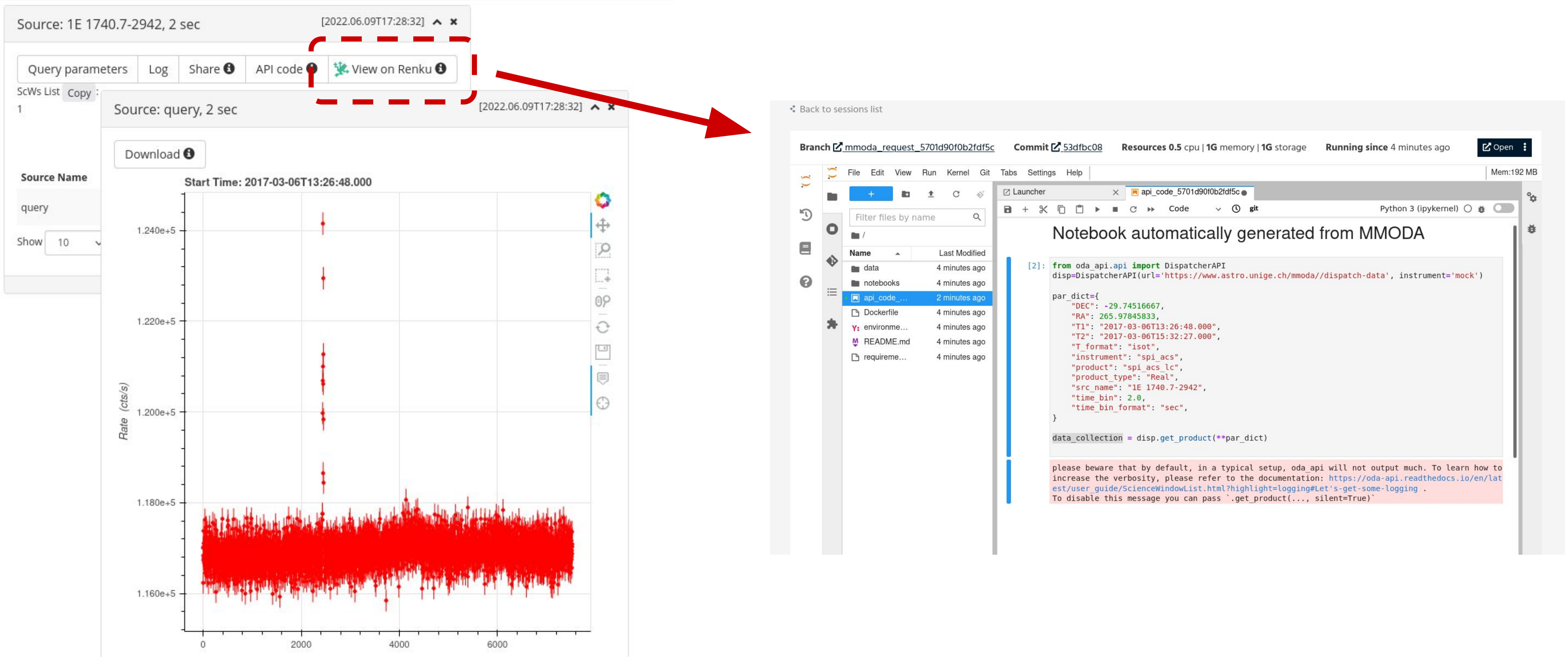
And even more so, we want to **leverage workflow as a service, possibly calling from another workflow**

We are publishing the live tool, not just it's output



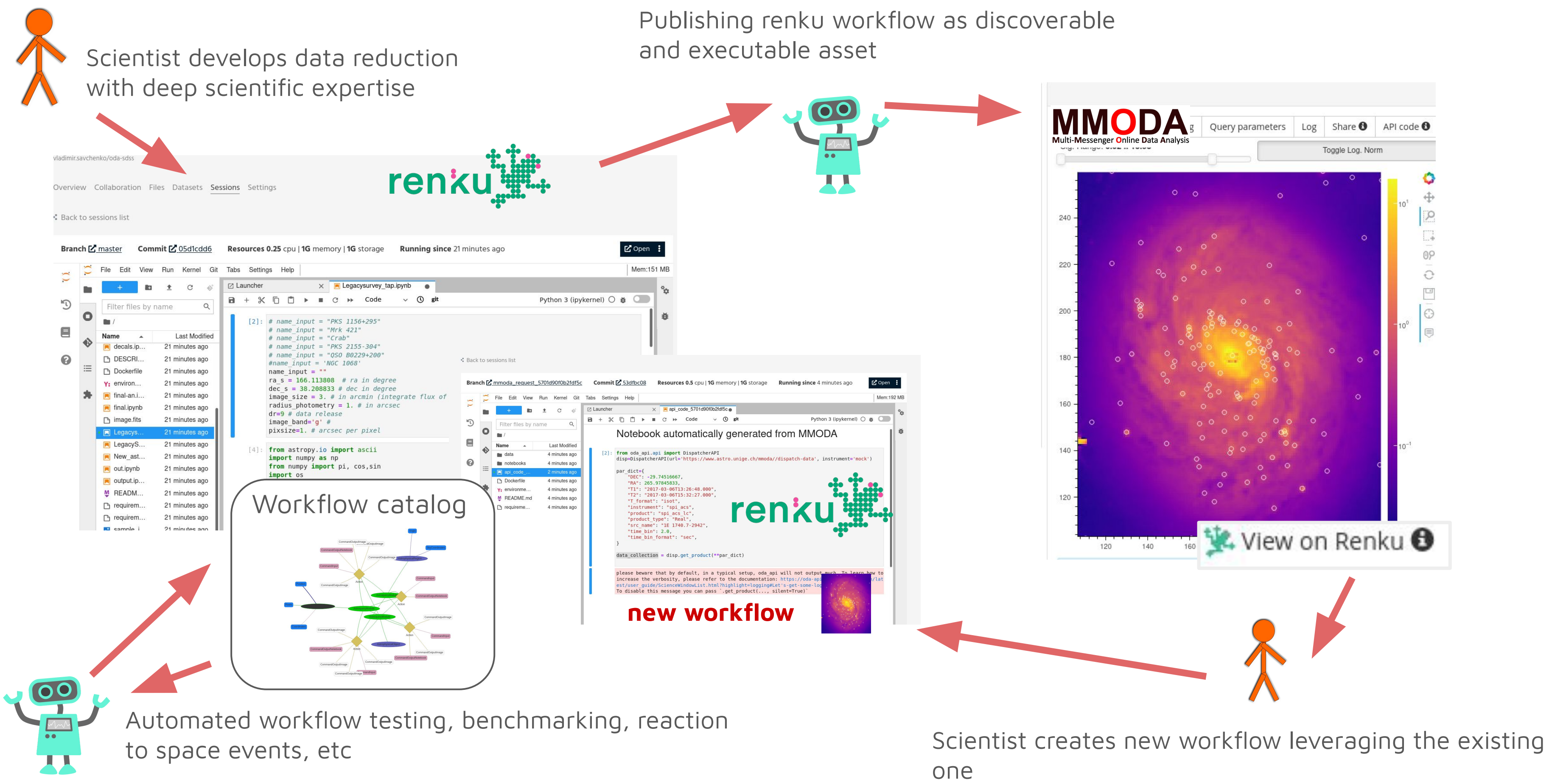
Helping to request MMODA services from Jupyter/Renku

Building new workflows by using results of the existing ones



Ontology-based service schema

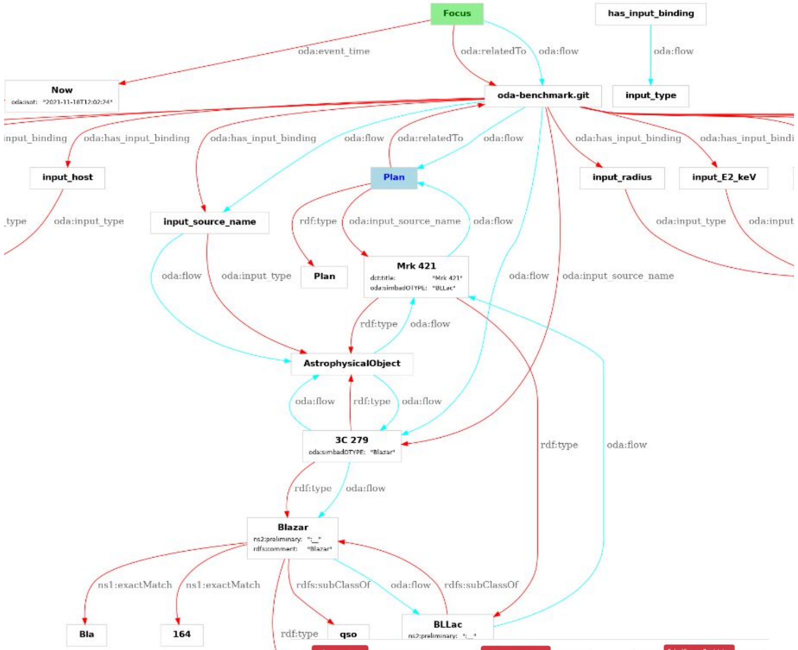
Feedback loop for crowd-sourcing workflow catalog



Add “creativity”: Linked Open Data Knowledge Graphs

People know a lot, and form free associations. Robots have much information too. E.g. much insight is reported in GCN Circulars but only accessible to people.

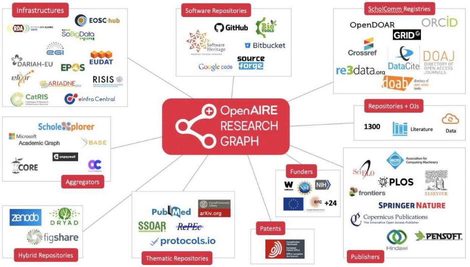
- **Global linked identifiers** - **URI** (ivo://, http://, ..): building common vocabularies. URIs point to documents, workflows, data, astro objects
- Explain possible **relations** in **ontologies**
- Embedding and following references, to express connections between different URIs

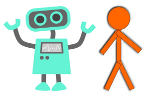


Need link more in the common KG language:

- we try to **consume graciously**
- for **annotating** and **publishing** integrate **code/workflows** with **data**: making sure we **produce cautiously**
- Using **graph relations** to rank and **optimize publication production**

Suitable balance between rigidity and flexibility of schemas. Linked Data ideas: make it connected into larger world with it’s language





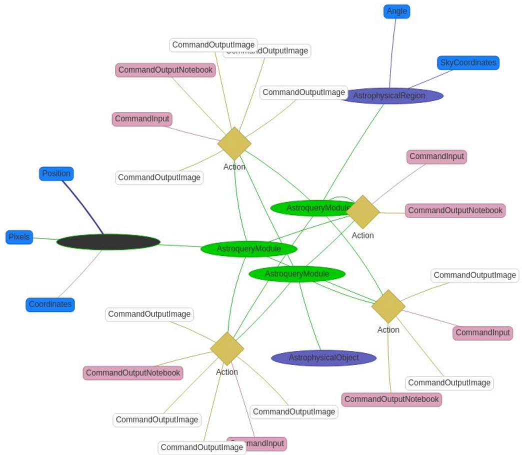
```
paper:gcn31435 paper:NUMBER "31435" ;
    paper:SUBJECT "GRB 220111A: BALROG localization …";
    paper:balrog_dec 6.380345e-01;
    paper:balrog_ra 1.498846e+02;
    paper:gbm_trigger_id 663621714 ;
    paper:grb_isot "2022-01-11T19:21:49.430000" ;
    paper:location <https://gcn.gsfc.nasa.gov/gcn3/31435.gcn3> ;
    paper:mentions_named_event <http://odahub.io/ontology/astroobject#GRB220111A>;
    paper:source "GCN" ;
    paper:url_json <https://grb.mpe.mpg.de/grb/GRB220111807/json> .
```

https://linked-open-data.space/ - small test example

Harvesting metadata for KG

To find these workflows in automation, we need suitable metadata.

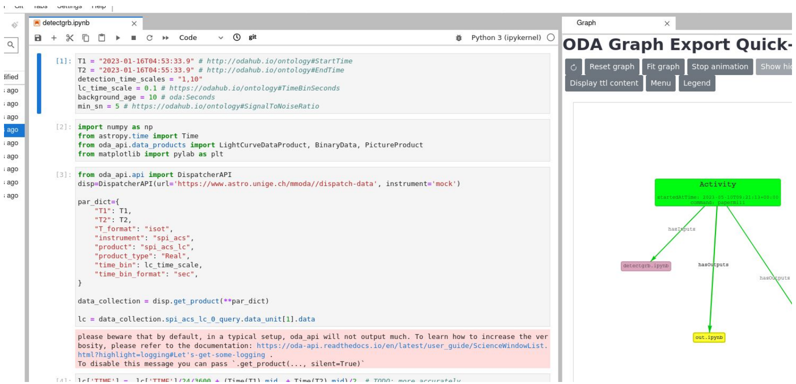
SDSC/Renku uses RDF KG for storing workflow metadata. We developed tools to **harvest astro-specific metadata from code runtime**, in particular by intercepting calls to other services/workflows, and the parameters used.



Interactive **Knowledge Graph** Explorer

To make the **KG more accessible to the user and the developer**, we made a jupyter plugin to show **provenance and ontology graphs** along within jupyterlab.

It helps to understand what was done and make choices for new annotations.



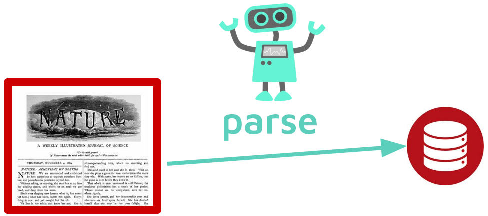
https://github.com/oda-hub/renku-aqs/

Time-domain astronomy domain has **micro-publications**: **short, rapid, high-impact, indexable/citable, with data embedded**

Everything shared can be equally considered a publication:

- new data segment
- automated alert
- nature paper

Robots can automatically harvest and parse them (e.g. to RDF)



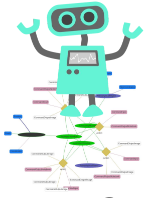
Need to build tools to explore and process these streams

URI	Title	Facts	Workflows
paper:gcn32182 2022-06-08T23:46:40	GRB 220609A: Fermi GBM Final Real-time Localization	DATE NUMBER SUBJECT gbm_dec gbm_ra gbm_rad gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_event mentions_named_event_type mentions_named_grb source	Explore
paper:gcn32178 2022-06-08T03:13:20	GRB 220608A: Fermi GBM Final Real-time Localization	DATE NUMBER SUBJECT gbm_dec gbm_ra gbm_rad gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_event mentions_named_event_type mentions_named_grb source	Explore
paper:gcn32174 2022-06-07T14:26:40	GRB 220606B: Fermi GBM Observation	DATE NUMBER SUBJECT cites cites_gcn_id gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_event mentions_named_event_type mentions_named_grb source	Explore
paper:gcn32181 2022-06-08T12:40:00	GRB 220606A: Fermi GBM observation	DATE NUMBER SUBJECT cites cites_gcn_id gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_event mentions_named_event_type mentions_named_grb source	Explore
paper:gcn32173 2022-06-07T11:40:00	GRB 220606B: Swift follow-up observations	DATE NUMBER SUBJECT cites cites_gcn_id gcn_authors gcn_from_email gcn_from_name mentions_grb mentions_named_event mentions_named_event_type mentions_named_grb reports_characteristic source	Explore

1. paper:DATE: ['22/06/07 13:36:38 GMT']
2. paper:NUMBER: ['32173']
3. paper:SUBJECT: ['GRB 220606B: Swift follow-up observations']
4. paper:cites: {'paper:gcn32170': {'@id': 'http://odahub.io/ontology/paper#gcn32170'}}
5. paper:cites_gcn_id: ['32170']
6. paper:gcn_authors: ['J.P. Osborne (U. Leicester), K.L. Page (U. Leicester), M.G. Bernardini (INAF-OAB), E. Ambrosi (INAF-IASFPA) , M. Capalbi (INAF-IASFPA), D.N. Burrows (PSU), J. D. Gropp (PSU), J.A. Kennea (PSU), P.A. Evans (U. Leicester), A. Breeveld (UCL-MSSL) report on behalf of the Swift team:']
7. paper:gcn_from_email: ['pae9@leicester.ac.uk']
8. paper:gcn_from_name: ['Phil Evans at U of Leicester']
9. paper:mentions_grb: ['body']
10. paper:mentions_grb_times: ['3']

https://github.com/oda-hub/literature-to-facts

KG-informed automation: **computational "Experiments"**



Since KG contains records of workflows, with I/O types, and data, it is easy to run "**experiments**": combine workflows with data and see what it gives.

***Processes that do compositions**, and **objective measures** are also registered workflows.*

- "**Act on new paper or observation report**"
- "**Act on new software or data**": re-do analysis of a "test case assumptions about instruments
- "**Act on new observation**": testing assumptions about physical reality
- "**Act on new platform or time moment**": make sure platform runs smoothly and is sane

View ▾ Act ▾ Paper ▾ Objectives ▾ Objects ▾ 0 more goals This service implements standards discussed here here and here. Source of the deployment and the application

3 from 0 to 3 days ago todays last week future past GCNs arXiv ATel Search for...

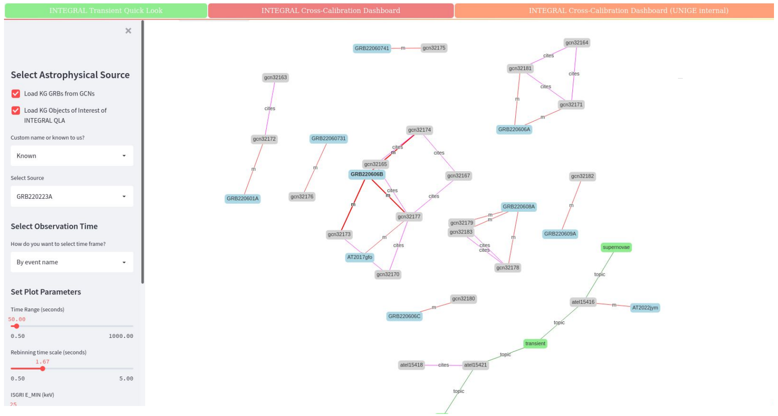
URI	Title	Facts	Workflows
paper:gcn29707 2021-03-24T11:24:43	GRB 210324A: Fermi GBM Final Real-time Localization	DATE NUMBER SUBJECT gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_grb source	Explore; paper origin — fetch-workflow-paper:gcn29707 — Run; INTEGRAL follow-up — Run; my affinity — oda:me oda:cares_for paper:gcn29707 — Care / Uncare
paper:gcn29715 2021-03-24T19:35:15	GRB 210324B: Fermi GBM Final Real-time Localization	DATE NUMBER SUBJECT gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_grb source	Explore
paper:gcn29718 2021-03-24T20:10:22	GRB 210324C: Fermi GBM Final Real-time Localization	DATE NUMBER SUBJECT gcn_authors gcn_from_email gcn_from_name grb_isot instrument mentions_grb mentions_named_grb source	Explore
paper:gcn29709	GRB 210323A: Fermi GBM detection		

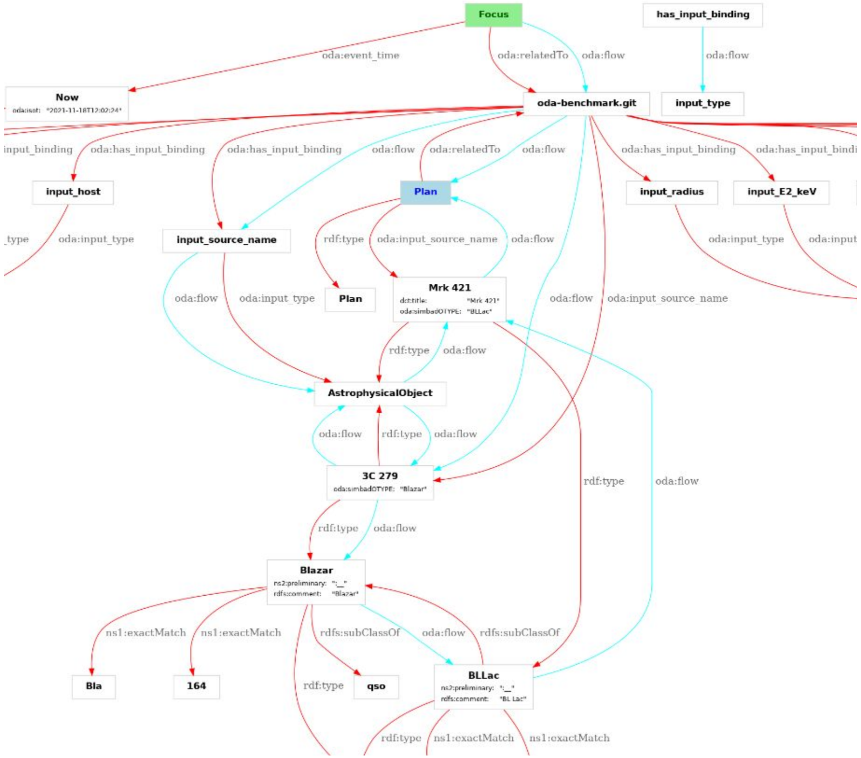
Building better papers by better leveraging information

Until now, we were building ad-hoc requests SPARQL to select and rank suitable workflows.

This does not represent our own intuition quite as much as we believe it is currently achievable.

We are trying to build some ML to optimize paper value (at minimum - impact)





http://odahub.io/ontology/

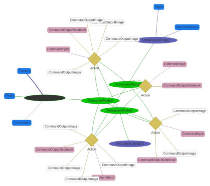
“Live” publications from data

To complete the science loop, we must rethink **production of publications** as well.

“Dead” publications kill the data by embedding mutilated irreproducible untraceable representations and references to it.

“Live” publications are themselves workflow outputs, compiled from data.

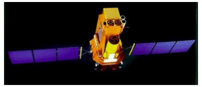
Linked Data



```
  Appendix~\ref{sec:association}).}

\begin{figure}
\centering
  \includegraphics[width=1.\linewidth]{acs_lc}
  \caption{SPI-ACS light curve of GRB~170817A (100~ms time
  resolution),
    detected 2 seconds after GW170817. The red line highlights the
    100~ms pulse. \ch{which has an S/N of
    \VAR{search.grb.snr|round(1)}} in
    SPI-ACS. The blue shaded region corresponds to a range of one
    standard deviation of the background.}
\label{fig:acslc}
\end{figure}
```

https://github.com/oda-hub/linked-data-latex/



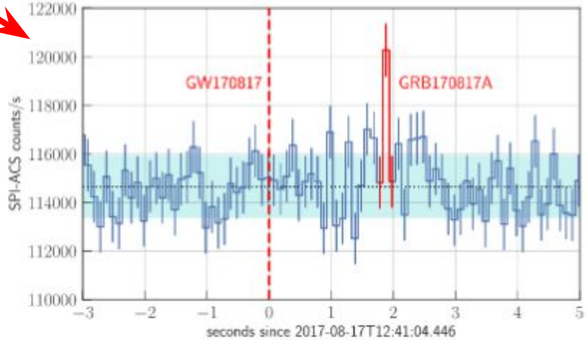


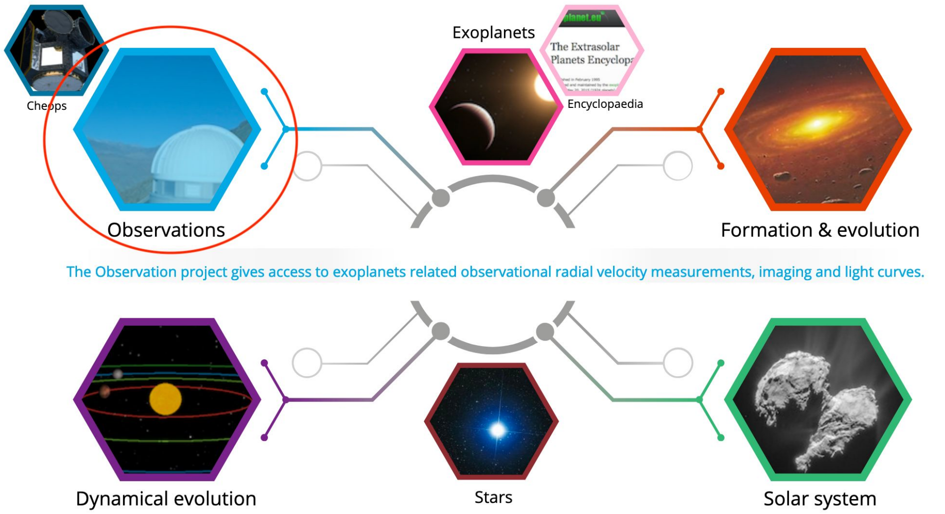
FIG. 2.— SPI-ACS light curve of GRB 170817A (100 ms time resolution), detected 2 seconds after GW170817. The red line highlights the 100 ms pulse, which has an S/N of 4.6 in SPI-ACS. The blue shaded region corresponds to a range of one standard deviation of the background.

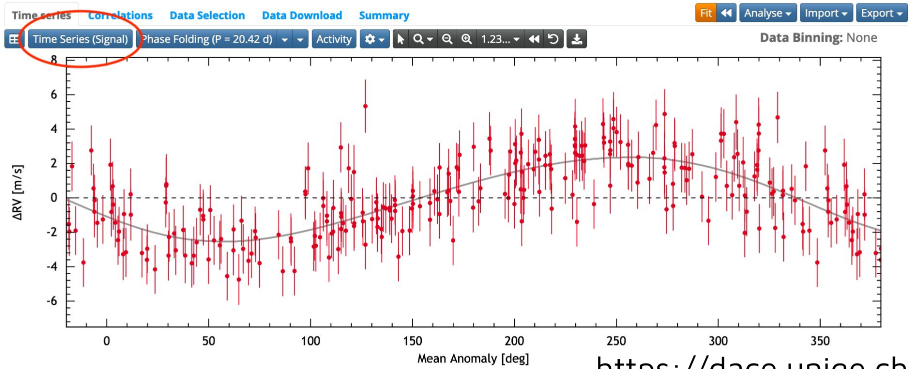
Synergies: DACE

Exoplanet community is facing similar challenges to that in the multi-messenger astronomy, and recently we discovered potential for synergy with **DACE project**, as **our goals and means started to overlap**

We are now actively developing horizontal integration and re-use of tools and technologies

Other communities finding that analysis services are beneficial, e.g. we are working with similar development in [Posydon]() project.

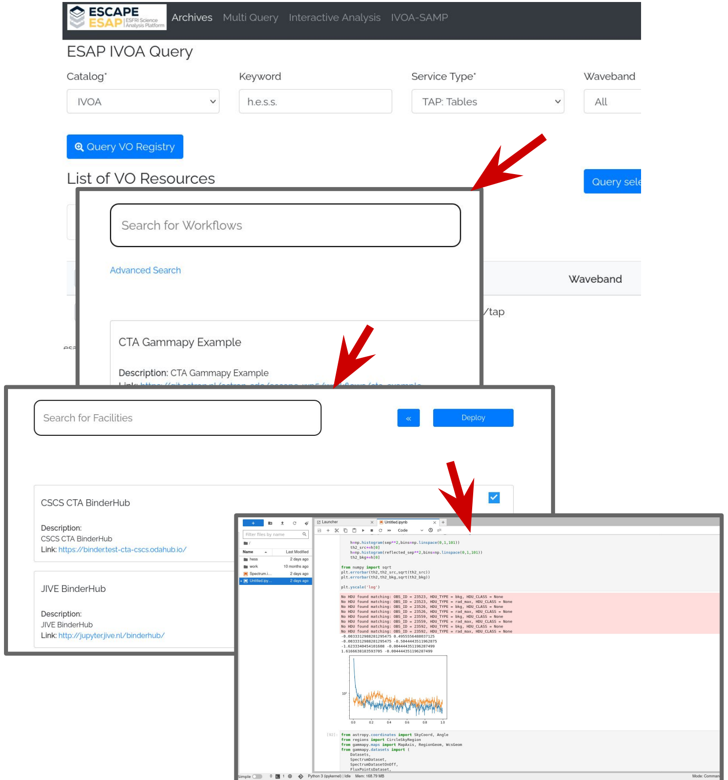




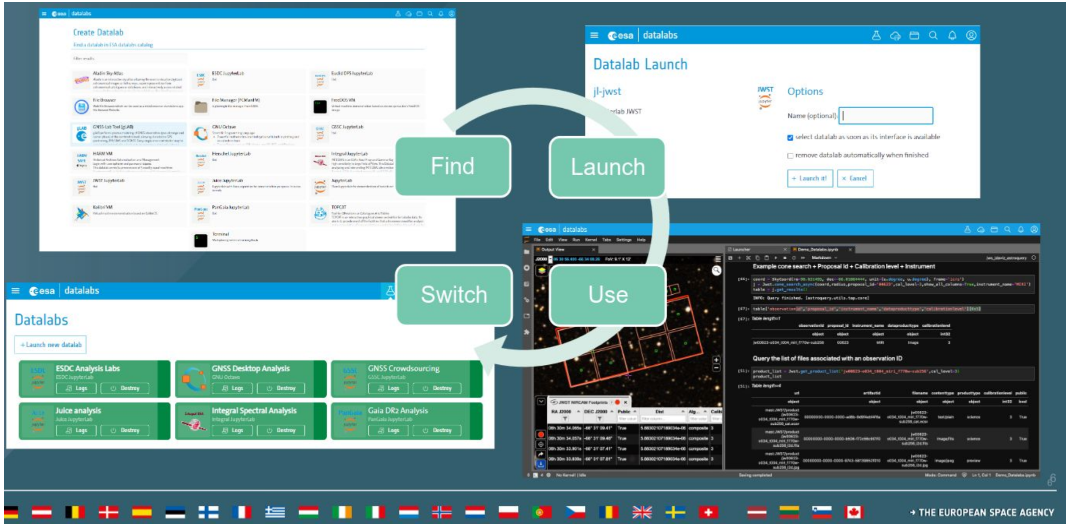
https://dace.unige.ch/

Synergies: Particle physics, ESAP/ESCAPE; ESA/DataLabs

EU/EOSC ESCAPE project developed **ESAP** platform joining particle physics and astronomy.



ESA developed DataLabs platform providing collection if diverse interactive environments for space science and technology. UNIGE/DepAstro was substantially involved in use case definition and design of its components and is now involved in platform integration.

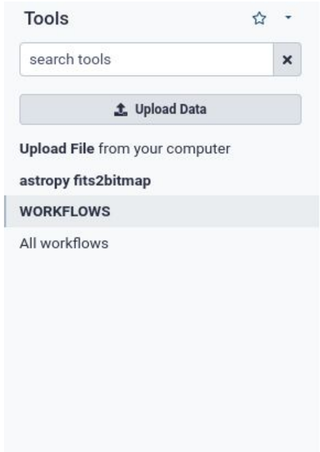


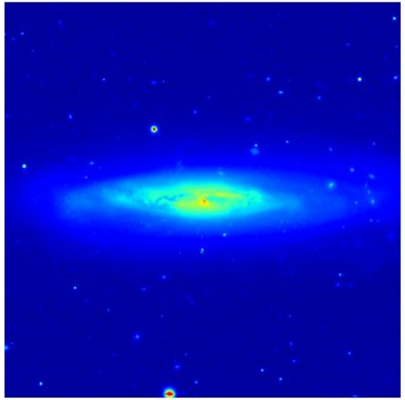
Synergies: EuroScienceGateway

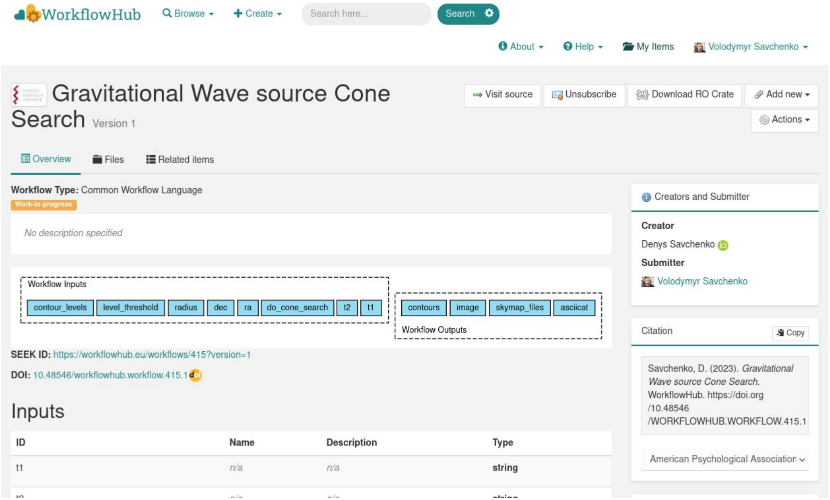
BioInformatics community has somewhat widely adapted **Galaxy platform** enabling open data analysis services with distributed compute.

We joined (along with particle physics, material science, climate science) an EOSC project to **bring our tools and workflows into Galaxy platform**.

The project also focuses on **integration with publications** and publishing workflows in WorkflowHub







Summary

Why
- **Rapidly growing astronomy** needs new **intelligent automation, both fast and creative** especially for:
  - **Multi-messenger and cross-domain** research
  - Ensuring science results **reuse and reproducibility**
  - Avoiding repetition and **saving energy**
  - **Teaching**

What
- Alongside with **Astronomical Data Centers** and often within, reusable practices for managing data analysis processes develop, leveraging **FAIR analysis "functions" (methods, workflows, tools) to Open Research Data.** Current project aims to **consolidate and establish** the **ecosystem of tools and technologies developed by us, leveraging synergies with other projects**
  - Focus on **workflow development environment** allows to **crowd-source tool creation** with our **MMODA** web-analysis platform
  - Workflow metadata in **KG** following **Linked Open Data** paradigm makes assets **discoverable** and **re-usable** fits well in **federalistic nature our community**

How
- **Sustainability** of this sort of activity is sometimes challenging, but the need to ensure it is recognized.
  - Continue **working closely with Data Centers** for their needs
  - **Added-value scientific services** beyond basic Data Centers (e.g. **multi-messenger**)
  - **Building up from infrastructure service provider level** (e.g. extra services along with a computing cluster)
  - Building **synergies (including technology integration)** with **other related projects** is one of the ways to make sure the developments continue being used and re-used. **In part for this, all our work is open**