UNIVERSITÉ
DE GENÈVE
FACULTÉ DES SCIENCES
Département d'astronomie

ISDC

RDF Knowledge Graph in Astronomy

our practical experience

Volodymyr Savchenko

UNIGE

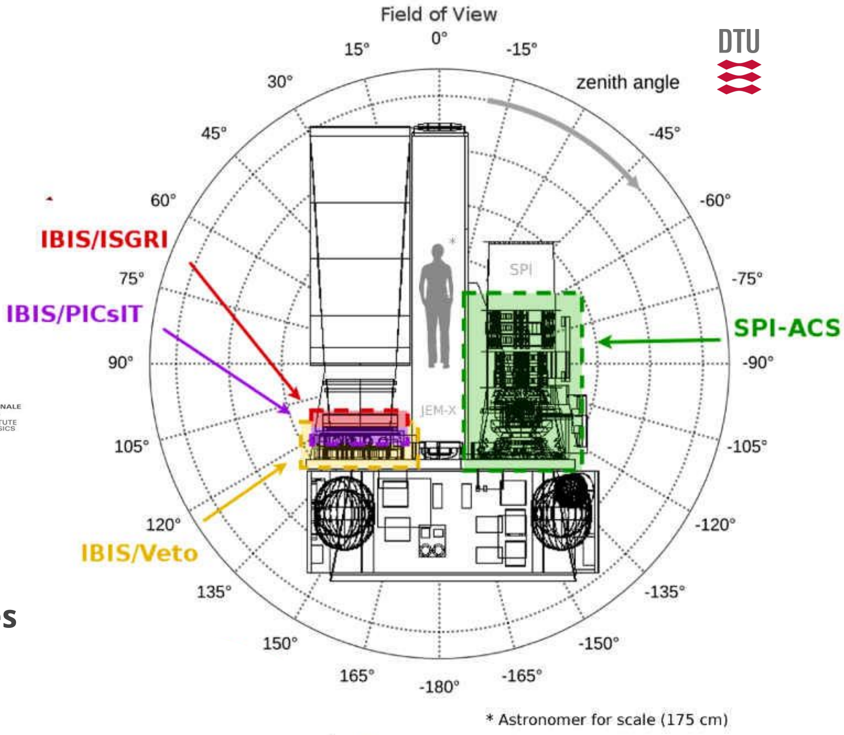
RDF Linked Data Meeting
23 - 04 - 2021

Overview

Purpose of my talk is to share evolution of our **domain goals**, **common approaches**, **solutions**, **future:**

- Context and **goals** in our **projects**
  - our role with INTEGRAL data center in Versoix
  - automated and human **workflows**
- Commonly adopted technologies
  - RDF in Astronomy
  - Move to cloud, ESA, EOSC
- Our own approach to automation and knowledge stewardship
- Future
  - Plans and hopes
  - open questions

INTEGRAL space observatory



* Astronomer for scale (175 cm)

cnrs

INAF ISTITUTO NAZIONALE DI ASTROFISICA NATIONAL INSTITUTE FOR ASTROPHYSICS

DTU

DLR

esa

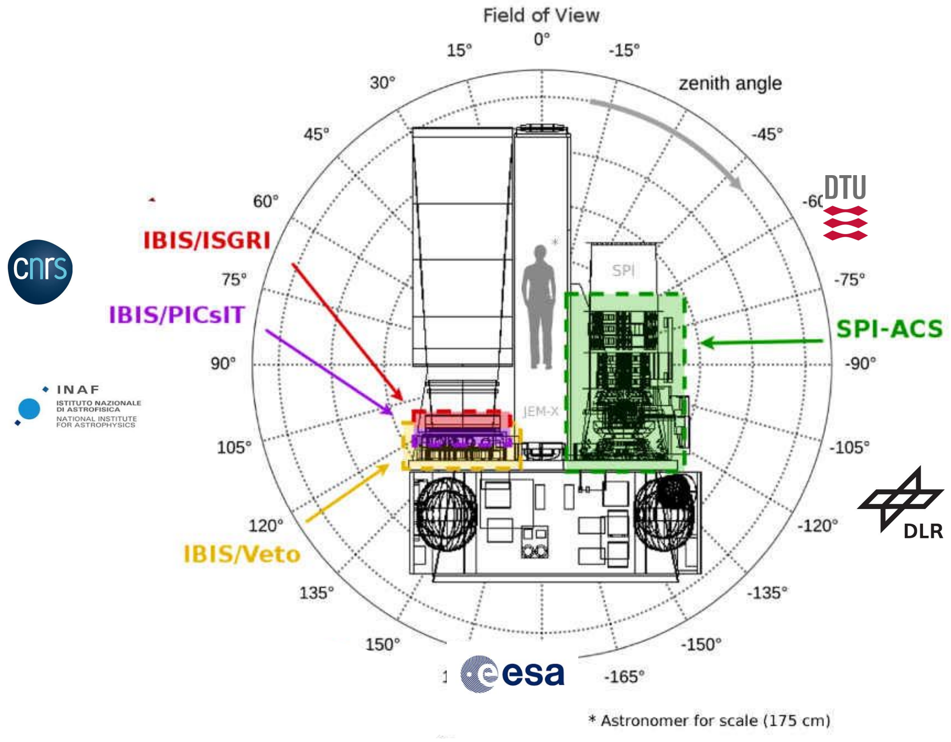
Best results can be achieved by **combining all INTEGRAL pieces**

But they are built by different countries, **challenging internal interoperability**

UNIVERSITÉ DE GENÈVE
FACULTÉ DES SCIENCES
Département d'astronomie

ISDC

INTEGRAL space observatory and context



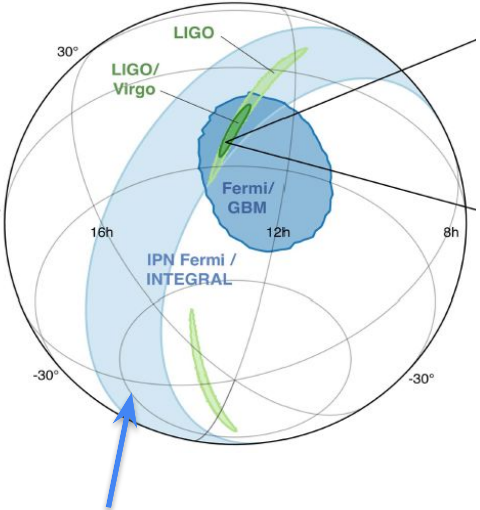
Core role of **ISDC** is to provide community with means to tackle the data: we **consolidate distribute software, services, data.**

In addition, INTEGRAL was not primarily built as a GRB detector, and provides best results by **combining with spectral, timing information from other missions**

Some INTEGRAL-only detections rarely get make impact on their own, it was decided to **provide API to access** public data, instead of sending uninformative **publications**.

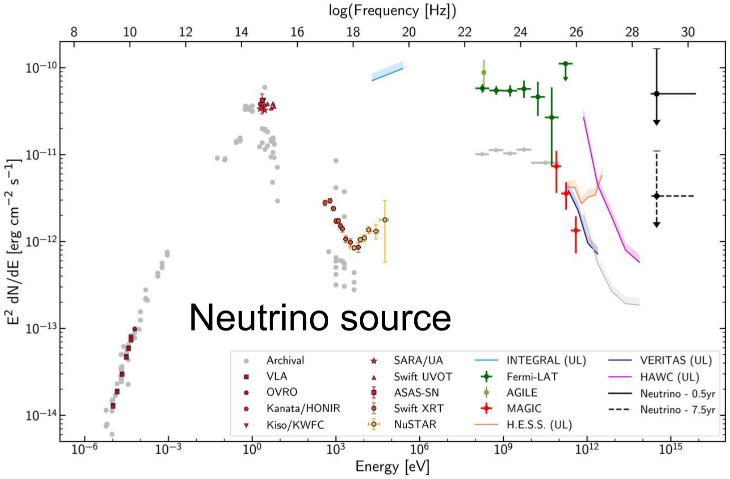
This mean that we need to **handle FAIR live services/APIs/workflows** not just **FAIR data**.

Multi-messenger astronomy is collaborative



Fermi + INTEGRAL Triangulation

unique multi-mission approach



Neutrino source

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** coincidence (2017)
- First detection of light emission from **high-energy neutrino source** (2018)
- Discovery of the source of mysterious **Fast Radio Burst** (2020)

VS+ 2017, LVC 2017

Traditional INTEGRAL analysis

Research, development environment lets experts develop, test, and integrate:

- data reduction
- theoretical models
- Spacecraft operation tools

Observers and Operators:

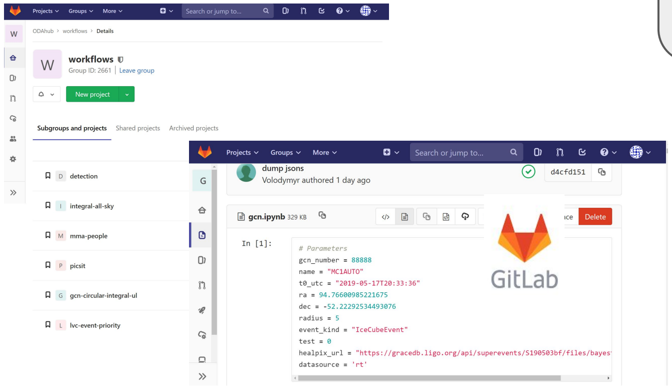
- **Find combinations** of data, adapters, statistical methods, publishers, planners
- **suggest new observations**
- **distribute** results in

ISDC

experts

data software documents

Observers, Operators, Shift



```
In [1]:
# Parameters
gcn_number = 88888
name = "MC1AUTO"
t0_utc = "2019-05-17T20:33:36"
ra = 94.76600985221675
dec = -52.22292534493076
radius = 5
event_kind = "IceCubeEvent"
test = 0
healpix_url = "https://gracedb.ligo.org/api/superevents/S190503bf/files/bayes
datasource = 'rt'
```

```
TITLE:   GCN CIRCULAR
NUMBER:  25505
SUBJECT: LIGO/Virgo S190828l:
in INTEGRAL SPI-ACS prompt ob
DATE:    19/08/28 08:59:07 GM
```

Nature events

VOEvent, HTTP, Kafka, etc

cnrs

INAF ISTITUTO NAZIONALE DI ASTROFISICA NATIONAL INSTITUTE FOR ASTROPHYSICS

esa

Challenges of our work (probably not too exceptional)

To facilitate this process, we need to:

- Discover and explore different **data**
- Access different **workflows** and use them on **data**
- Be **informed** but also conscious of **information veracity**
- Try many combinations **quickly**: be exhaustive and ready for unexpected
- Setup automation for **low latency** and **high-volume activity**

None of this is new for many people, and solutions to these challenges were implemented before, in Astronomy not the least of all.

We are coordinating with other Astrophysical Mission Science Data Centers and space agencies on developing common solutions to these issues.

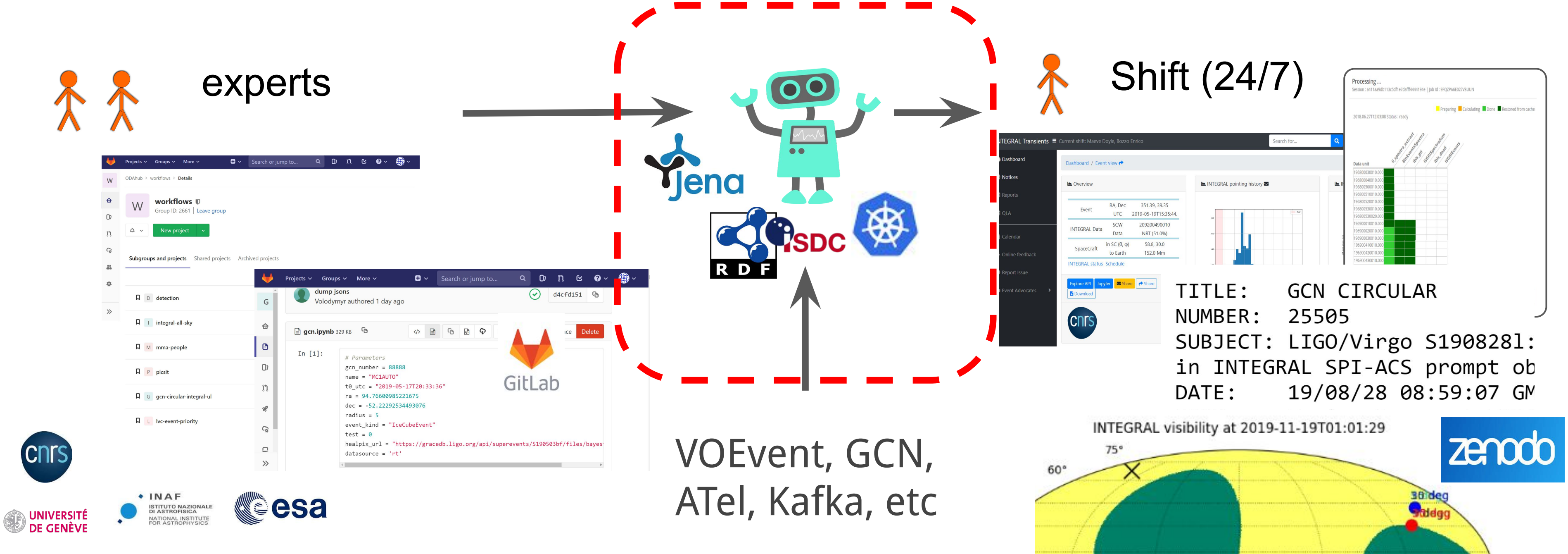
We always search for **new ways to tackle** these challenges, especially when development from scratch is limited!

“Standard” modern INTEGRAL transient analysis

Research, development environment lets experts develop, test, and integrate:

- data reduction (close to data)
- theoretical models (linked to literature)
- statistical methods (as portable as possible)
- Spacecraft operation tools

- **Find combinations** of data, adapters, statistical methods, publishers, planners
- **suggest follow-up**
- **distribute** standard results with public data, uploads to zenodo sandbox.



```
In [1]:
# Parameters
gcn_number = 88888
name = "MC1AUTO"
t0_utc = "2019-05-17T20:33:36"
ra = 94.76600985221675
dec = -52.22292534493076
radius = 5
event_kind = "IceCubeEvent"
test = 0
healpix_url = "https://gracedb.ligo.org/api/superevents/S190503bf/files/bayes
datasource = 'rt'
```

Event	RA, Dec	351.39, 39.35
	UTC	2019-05-19T15:35:44.
INTEGRAL Data	SCW	209200490010
	Data	NRT (51.0%)
SpaceCraft	in SC (θ, φ)	58.8, 30.0
	to Earth	152.0 Mm

INTEGRAL status Schedule

```
TITLE:    GCN CIRCULAR
NUMBER:   25505
SUBJECT:  LIGO/Virgo S190828l:
in INTEGRAL SPI-ACS prompt ob
DATE:     19/08/28 08:59:07 GM
```

INTEGRAL visibility at 2019-11-19T01:01:29

VOEvent, GCN, ATel, Kafka, etc

approaches

Addressing our challenges relies on **FAIR** Work Space

To support this environment, we need robust means of **Finding, Accessing, Interoperating, and Reusing** our assets:

- Documents
- Data
- **Workflows**
  - **Code: scientific software which we build and distribute**
  - **API's: Software as a Service, which we provide**
  - **Humans**

Publishing, communication is and always has been at the **core of scientific activity**. But means of publishing go beyond **paper manuscripts**.

Now, we also need to **publish**, share our **data**, **code**, **services**, and **integrate** different means of publishing.

Especially interesting aspect is in sharing a diversity of:

- Workflows
    - Code: **scientific software which we build and distribute**
        - Official software (source, binary, doc, recently containers)
        - Helper scripts and packages: gitlab, github
        - INTEGRAL **Quick Look Analysis** results
    - Web-based Software: **API's and Frontends, which we provide to the community**
        - **Data browse** interface (developed by NASA, hosted un Versoix, since ~2002)
        - Help desk, **issue** handling and resolution
        - **Realtime data** interoperability (since 2011)
        - **AstroODA** online analysis (internally since before 2017, first public in 2019)
            - And all of it's **backends** separately
        - **Multi-Messenger Transient Analysis** (since 2018)
        - Various smaller API's for specific purposes
    - Humans: **tools and technologies for managing ourselves**
        - Support of the software and services
        - Expertise in scientific instrumentation and data analysis methods

Sharing workflows meaningfully (in a **FAIR** way) is especially hard!

These are just ours, we also want to find and leverage external ones (sometime by wrapping in ours).

Addressing our challenges relies on **FAIR** Work Space

Astronomy is quite aware of **FAIR** challenges.

International Virtual Observatory Alliance (**IVOA**) made a very substantial effort to promote FAIR practices in Astronomy, much of it based on **RDF**/**Linked Data**, including recent large projects like ASTERICS:

https://www.asterics2020.eu/

Provenance is widely recognized in Astro community and adopted as key feature in some large community projects.

https://www.ivoa.net/documents/ProvenanceDM/

RDF Ontologies for sharing were developed, in different degrees of readiness:

https://www.ivoa.net/rdf/

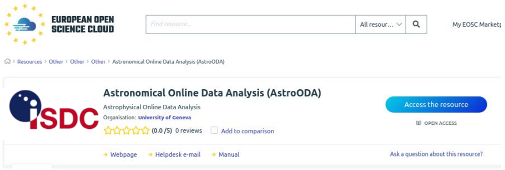
Centre de Données astronomiques de Strasbourg (CDS) has developed a variety of taxonomies and ontologies for astronomy

http://cdsweb.u-strasbg.fr/

These developments provide great basis for describing data, but not so much processes/workflows/APIs

Addressing our challenges relies on **FAIR** Work Space

EOSC allows to publish and annotate service offerings, helping **live service interoperability.**



Astronomical Online Data Analysis (AstroODA)

Astrophysical Online Data Analysis

Organisation: University of Geneva

☆☆☆☆☆ (0.0 /5) 0 reviews

EUROPEAN OPEN SCIENCE CLOUD
Find resource...
All resour...
My EOSC Market
Resources › Other › Other › Other › Astronomical Online Data Analysis (AstroODA)
iSDC
□ Add to comparison
Access the resource
OPEN ACCESS
Webpage
Helpdesk e-mail
Manual
Ask a question about this resource?

ESA currently actively develops similar approach to scientific asset stewardship: **ESA DataLabs** (and we are involved in it as early adopters). They provide a platform to handle Astronomical workflows and are actively working on a **DataLab (~workflow) discovery hub.**

Renku of **SDSC** addresses some of the keys needs valuable for further developments, and has a **Knowledge Graph**, so we could interoperate with it.

So we decided to improve our **Project Knowledge Base** by adding an **RDF Knowledge Graph** to glue together our **very heterogeneous assets with complex constraints**.

ODA Knowledge Base (and Graph) - since 2018

Store all references to data and workflows in KG.

Store **bigger data in buckets, serial in tables**. Explain in KG how to access them.

OWL2 and **SHACL** when possible. But also not always over-obsess over definitions, turns out sometimes ontology can be defined and enforced as KG evolves.

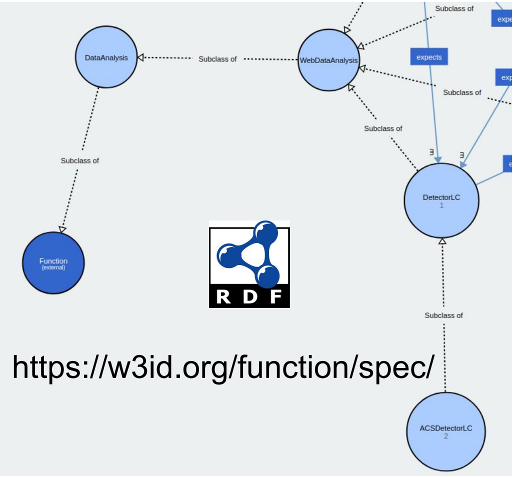
Ontology of **processes** based on **fno** when feasible: most important to define input and output formats (types), this allows composition.

Simple domain-specific **literature parsing** ingests new events in the KG.

Documents, **analysis results**, especially **workflow** executions, ingested and annotated.

Scientific terms and concepts and **relations** between are stored, as much as needed by our case: a lot of specific things, some general with common terms (e.g. Crab_pulsar is a neutron star) when possible.

Human actors are explicit focus of the KG.



https://w3id.org/function/spec/

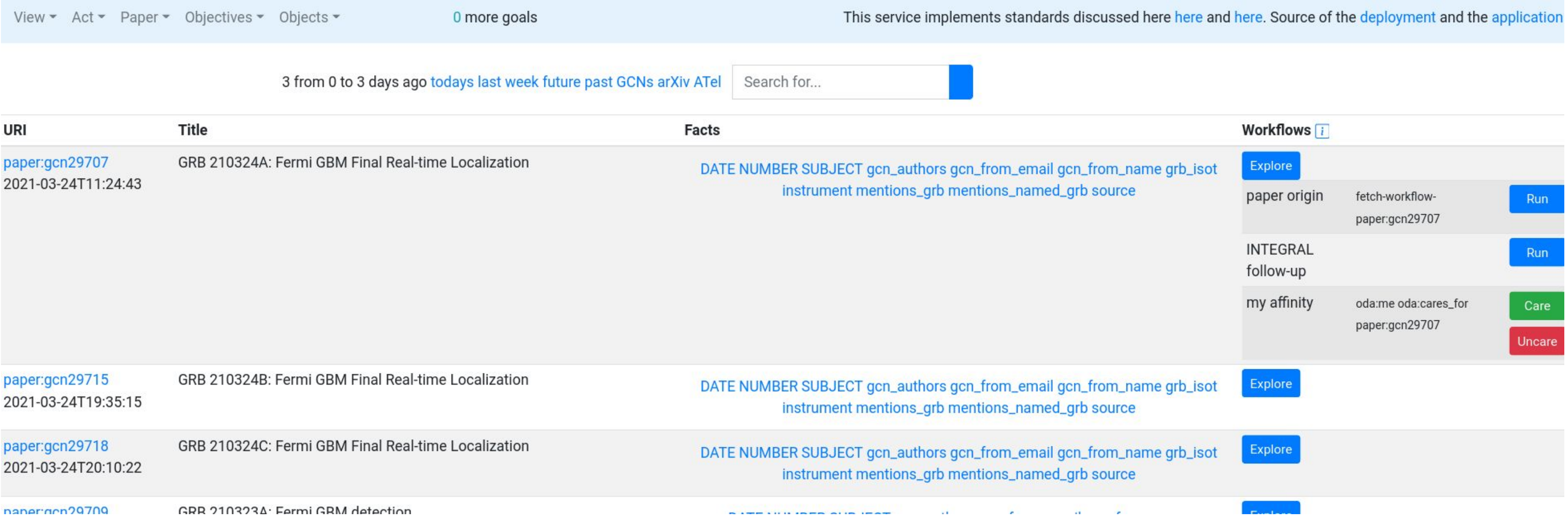
ODA Computational “Experiments”

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.*

Real, emergent, classes of compositions:

- “**Act on new paper or observation report**”
- “**Act on new software or data**”: re-do analysis of a “test case”, ensuring 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



0 more goals

This service implements standards discussed here here and here. Source of the deployment and the application

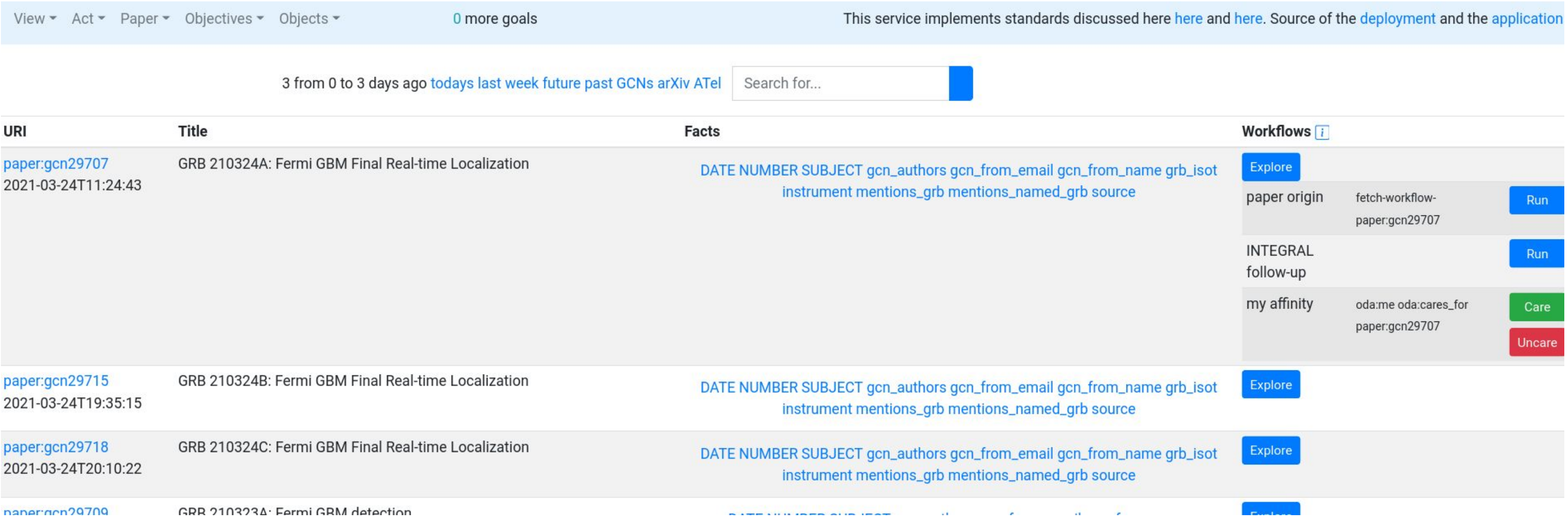
3 from 0 to 3 days ago todays last week future past GCNs arXiv ATel

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	DATE NUMBER SUBJECT	Explore

ODA Computational “Experiments”

Composition is substituting parameters in workflows, it builds workflows from workflows, and hence is a form of “reasoning” on the graph.

Workflow **executions** are also “reasoning” since they transform non-trivial workflows (e.g. python functions, or jupyter notebooks) into data-fetching workflows.



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	DATE NUMBER SUBJECT gcn_authors gcn_from_email gcn_from_name	Explore

Plans and Future needs

Obviously our “**KG**” is **very simple and goal-driven**, and hence quite specific to our case.

- As we realized before, **some of the typical astronomy activity can be automated-out.** How much more can be? Service operations experience shows scientist ask predictable things.
- We did not explore any but simple workflow compositions! There is a lot of potential and interest from our domain partners.
- We are continuously developing it and it might be good to change it to another framework if needed. Open for exploring options!
- Would be good to handle **ontologies** better, much new ones develop in advance?
- Interoperability of graphs? More user engagement? Publish it publicly?
- We have little graph analysis intelligence so far, even if a lot of “reasoning” by workflows operating on graph.

Problems, Questions

- Performance is a challenge. For about 1M triples it is just about ok.
  - we store data outside, e.g. in buckets, and query with KG-defined workflows
  - multiple graphs and jena instances are used. Some are small and public. Some have all kinds of unprocessed inputs
  - Sometimes we make simpler SPARQL query and do the rest in rdflib, locally
- How to set fine-grained permissions for many users?
  - we make several graphs, but that’s not fine-grained enough
- Developing ontology is hard. More tools are needed
  - we relied on “just ingesting” and then frequent refactoring, by specifying reasoning rules. It becomes part of the natural graph evolution. But more tools are needed to support this workflow.
- Defining what is authoritative claim
  - Who has a right to make propositions/claims? There are mechanisms with named graphs. We also tried RDF*.
  - Atomic Data restriction on RDF can be useful
- It is not guaranteed that KG itself is reproducible, it’s hard to make it append-only, with versioning
  - We adopted a form of reification for this, but it is complex
- All these mitigations make the **KG very verbose.**
  - Trying to use different graph views and subset to deal with it