Enriching a Time-Domain Astrophysics Corpus with Named Entity, Coreference, and Astrophysical Relationship Annotations

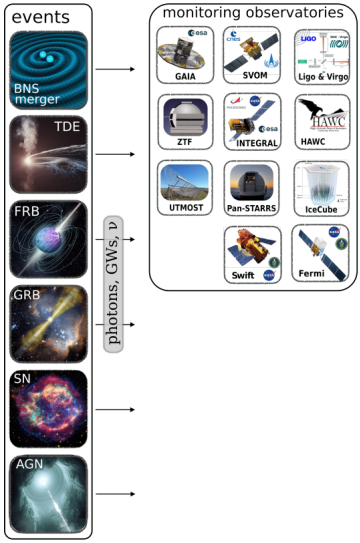
Atilla Kaan Alkan[1,2],
Felix Grezes[3], Cyril Grouin[1], Fabian Schüssler[2] and Pierre Zweigenbaum[1]

(1) Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405 Orsay, France
(2) IRFU, CEA, Université Paris-Saclay, F-91191 Gif-sur-Yvette, France
(3) Harvard-Smithsonian Center for Astrophysics, 60 Garden Street, Cambridge, MA 02138, USA

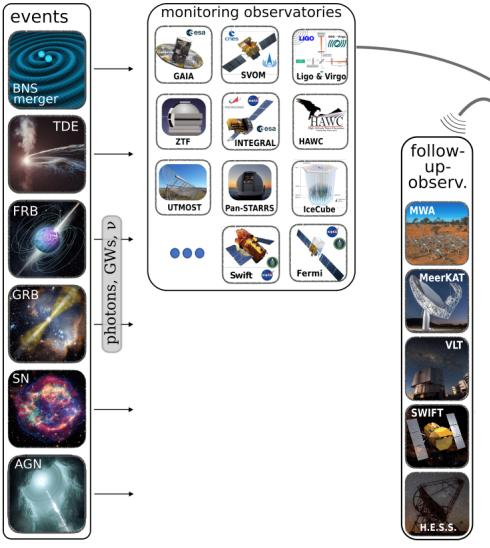
LREC-COLING 2024, Torino, Italy

Content

Astronomical Observations



Astronomical Observations



Schüssler, **PR+** (2021, *ICRC*)
PR+ (2021, *ADASS*)
PR+ (2021, *ApJS*)

Human-written messages

Fermi-LAT gamma-ray observations of IceCube-220918A and detection of a new gamma-ray source, Fermi J0502.5+0037

ATel #15620; *S. Garrappa (DESY-Zeuthen), S. Buson (Univ. of Wuerzburg) and J. Sinapius (DESY-Zeuthen) on behalf of the Fermi-LAT collaboration:*
*on **19 Sep 2022; 20:50 UT***
Credential Certification: Simone Garrappa (simone.garrappa@gmail.com)

Subjects: Gamma Ray, Neutrinos, Request for Observations, AGN, Blazar

TITLE: GCN CIRCULAR
NUMBER: 32562
SUBJECT: IceCube--220918A - IceCube observation
DATE: 22/09/18 14:54:14 GMT
FROM: Erik Blaufuss at U. Maryland/IceCube

The IceCube Collaboration (http://icecube.wisc

On 2022-09-18 at 12:46:05.32 UT IceCube detect
event was selected by the ICECUBE_Astrotrack_B
This alert has an estimated false alarm rate o
operating state at the time of detection.

After the initial automated alert (https://gcn
algorithms have been applied offline, with the direction refined to:

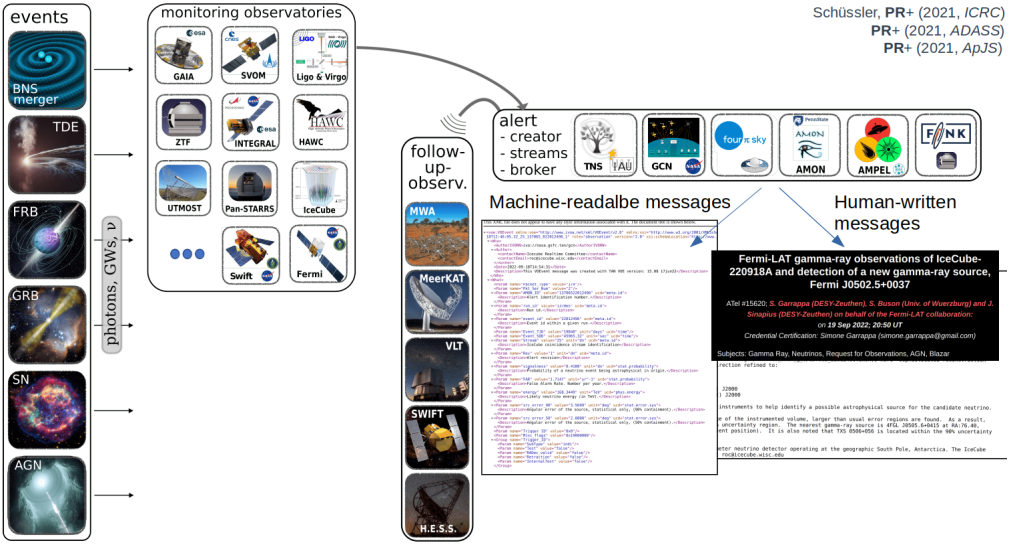
Date: 2022-09-18
Time: 12:46:05.32 UT
RA: 75.64 +3.79 / -3.38 deg 90% PSF containment) J2000
Dec: +3.58 (+3.79 / -3.40 deg 90% PSF containment) J2000

We encourage follow-up by ground and space-based instruments to help identify a possible astrophysical source for the candidate neutrino.

Given the geometry of the event, skimming the edge of the instrumented volume, larger than usual error regions are found. As a result, several Fermi 4FGL or 3FHL sources are in the 90% uncertainty region. The nearest gamma-ray source is 4FGL J0505.6+0415 at RA:76.40, Dec: 4.27 deg (1.42 deg away from the best-fit event position). It is also noted that TXS 0506+056 is located within the 90% uncertainty region.

The IceCube Neutrino Observatory is a cubic-kilometer neutrino detector operating at the geographic South Pole, Antarctica. The IceCube realtime alert point of contact can be reached at roc@icecube.wisc.edu

Astronomical Observations



Schüssler, **PR+** (2021, *ICRC*)
PR+ (2021, *ADASS*)
PR+ (2021, *ApJS*)

Fermi-LAT gamma-ray observations of IceCube-220918A and detection of a new gamma-ray source, Fermi J0502.5+0037

ATel #15620; *S. Garrappa (DESY-Zeuthen), S. Buson (Univ. of Wuerzburg) and J. Sinapius (DESY-Zeuthen) on behalf of the Fermi-LAT collaboration:*
*on **19 Sep 2022; 20:50 UT***
Credential Certification: Simone Garrappa (simone.garrappa@gmail.com)

Subjects: Gamma Ray, Neutrinos, Request for Observations, AGN, Blazar

Towards an Automated Analysis of Observation Reports

The Chandra X-ray Observatory imaged the field of GRB 991216 for 10ks starting on Dec 18.208 (UT), i.e. about 37 hrs after the GRB (Kippen et al. GCN n. 463). From a preliminary analysis of the data we found a single unknown point source, that we name CXO J050931.4+111706, in the center of the field. The position is RA = 05h09m31.35s, DEC = 11d17'.

Celestial Object	Observatory	Date	Coordinates	Citation
CXO J050931.4+111706	Chandra X-ray Observatory	Dec 18.208 (UT)	RA = 05h09m31.35s, DEC = 11d17'	Kippen et al. GCN n. 463
GRB 991216	...	...	...	...

Annotation Obstacles

- **Obstacles**:
  - **Availability:**
    - ABC, Astro are annotated corpora but not available;
    - The DEAL Shared Task and TDAC corpus are the only available corpora.
  - **Annotation diversity:**
    - Type of documents: mainly scholarly papers;
    - Type of annotation: only NER-oriented corpora $\rightarrow$ not enough to follow fine-grained information in text.

Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.

MASTER OT J105440.86-391319.0 - PSN in PGC600519 (2.9"E,7"N) MASTER-SAAO auto-detection system (Lipunov et al. , " MASTER Global Robotic Net " , Advances in Astronomy , 2010 , 30L) discovered OT source at (RA , Dec) = 10h 54 m 40.86s , -39d 13 m 19s.0 on 2016 - 10 - 08 02:56:46.019UT with m_OT=16.5 . This PSN is in 2.9"E,7"N from the center of PGC600519 and is seen on 8 images . We have reference image without OT on 2015 - 02 - 27 22:35:30UT with m_lim=19.8 Spectral observations are required The discovery and reference images are http://master.sai.msu.ru/static/OT/MASTEROTJ105440.86-391319.0.jpg. MASTER OT J160918.30 - 333456.7 discovery - bright possible dwarf nova outburst , ampl>5.4 m MASTER-SAAO auto-detection system discovered OT source at (RA , Dec) = 16h 09 m 18.30s -33d 34 m 56.7s on 2016 - 09 - 21.72860 UT . The OT unfiltered magnitude is 16.6 m (mlim=18.6 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 04 - 04.98706 UT with unfiltered magnitude limit 19.7 m. There is only GSC2.3.2 known star with known only blue jmag=18.11(1975year - previous outburst) The 22 m POSS limit gives the amplitude of current outburst & gt;5.4 m Spectral observations are required . The discovery and reference images are available at : http://master.sai.msu.ru/static/OT/160918.30-333456.7.png. MASTER OT J155546.00 - 734455.8 discovery - OT with ampl>4.6 m , no VIZIER data MASTER-SAAO auto-detection system discovered OT source at (RA , Dec) = 15h 55 m 46.00s -73d 44 m 55.8s on 2016 - 09 - 19.83295 UT . The OT unfiltered magnitude is 17.4 m (limit 18.5 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 03 - 09.84253 UT with unfiltered magnitude limit 20.8m . There is no any sources in VIZIER database inside . Spectral observations are required . The discovery and reference images are available at : http://master.sai.msu.ru/static/OT/155546.00 - 734455.8.png.

Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.

Celestial Object

MASTER OT J105440.86-391319.0 - PSN in PGC600519 (2.9"E,7"N) MASTER-SAAO auto-detection system (Lipunov et al. , " MASTER Global Robotic Net " , Advances in Astronomy , 2010 , 30L) discovered OT source at (RA , Dec) = 10h 54 m 40.86s , -39d 13 m 19s.0 on 2016 - 10 - 08 02:56:46.019UT with m_OT=16.5 . This PSN is in 2.9"E,7"N from the center of PGC600519 and is seen on 8 images . We have reference image without OT on 2015 - 02 - 27 22:35:30UT with m_lim=19.8 Spectral observations are required The discovery and reference images are

Celestial Object

http://master.sai.msu.ru/static/OT/MASTEROTJ105440.86-391319.0.jpg. MASTER OT J160918.30 - 333456.7 discovery - bright possible dwarf nova outburst , ampl>5.4 m MASTER-SAAO auto-detection system discovered OT source at (RA , Dec) = 16h 09 m 18.30s -33d 34 m 56.7s on 2016 - 09 - 21.72860 UT . The OT unfiltered magnitude is 16.6 m (mlim=18.6 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 04 - 04.98706 UT with unfiltered magnitude limit 19.7 m. There is only GSC2.3.2 known star with known only blue jmag=18.11(1975year - previous outburst) The 22 m POSS limit gives the amplitude of current outburst & gt;5.4 m Spectral observations are required . The discovery and reference images are

Celestial Object

available at : http://master.sai.msu.ru/static/OT/160918.30-333456.7.png. MASTER OT J155546.00 - 734455.8 discovery - OT with ampl>4.6 m , no VIZIER data MASTER-SAAO auto-detection system discovered OT source at (RA , Dec) = 15h 55 m 46.00s -73d 44 m 55.8s on 2016 - 09 - 19.83295 UT . The OT unfiltered magnitude is 17.4 m (limit 18.5 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 03 - 09.84253 UT with unfiltered magnitude limit 20.8m . There is no any sources in VIZIER database inside . Spectral observations are required . The discovery and reference images are available at : http://master.sai.msu.ru/static/OT/155546.00 - 734455.8.png.

Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.

Celestial Object
MASTER OT J105440.86-391319.0 - PSN in PGC600519 (2.9"E,7"N) MASTER-SAAO auto-detection system (Lipunov et al. , " MASTER Global Robotic Net " , Advances in Astronomy , 2010 , 30L)

Celestial Region Date
discovered OT source at (RA , Dec) = 10h 54 m 40.86s , -39d 13 m 19s.0 on 2016 - 10 - 08 02:56:46.019UT with m_OT=16.5 . This PSN is in 2.9"E,7"N from the center of PGC600519 and is seen on 8 images . We have reference image without OT on 2015 - 02 - 27 22:35:30UT with m_lim=19.8 Spectral observations are required The discovery and reference images are

Celestial Object
http://master.sai.msu.ru/static/OT/MASTEROTJ105440.86-391319.0.jpg. MASTER OT J160918.30 - 333456.7 discovery - bright possible dwarf nova outburst , ampl>5.4 m MASTER-SAAO

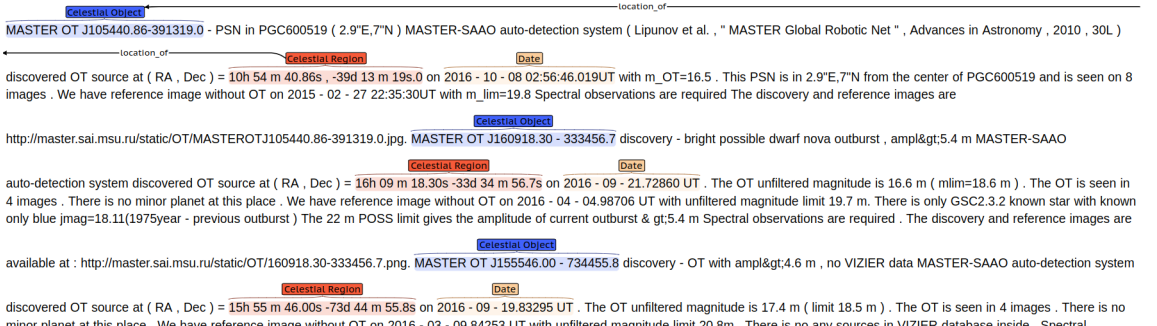
Celestial Region Date
auto-detection system discovered OT source at (RA , Dec) = 16h 09 m 18.30s -33d 34 m 56.7s on 2016 - 09 - 21.72860 UT . The OT unfiltered magnitude is 16.6 m (mlim=18.6 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 04 - 04.98706 UT with unfiltered magnitude limit 19.7 m. There is only GSC2.3.2 known star with known only blue jmag=18.11(1975year - previous outburst) The 22 m POSS limit gives the amplitude of current outburst & gt;5.4 m Spectral observations are required . The discovery and reference images are

Celestial Object
available at : http://master.sai.msu.ru/static/OT/160918.30-333456.7.png. MASTER OT J155546.00 - 734455.8 discovery - OT with ampl>4.6 m , no VIZIER data MASTER-SAAO auto-detection system

Celestial Region Date
discovered OT source at (RA , Dec) = 15h 55 m 46.00s -73d 44 m 55.8s on 2016 - 09 - 19.83295 UT . The OT unfiltered magnitude is 17.4 m (limit 18.5 m) . The OT is seen in 4 images . There is no minor planet at this place . We have reference image without OT on 2016 - 03 - 09.84253 UT with unfiltered magnitude limit 20.8m . There is no any sources in VIZIER database inside . Spectral observations are required . The discovery and reference images are available at : http://master.sai.msu.ru/static/OT/155546.00 - 734455.8.png.

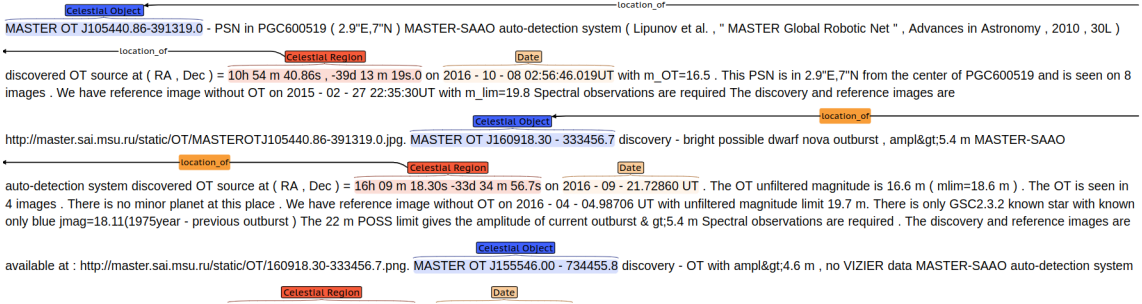
Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.



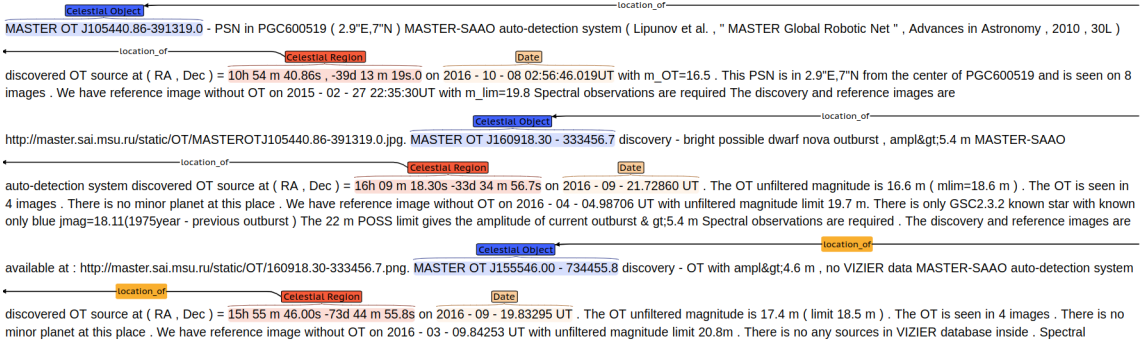
Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.



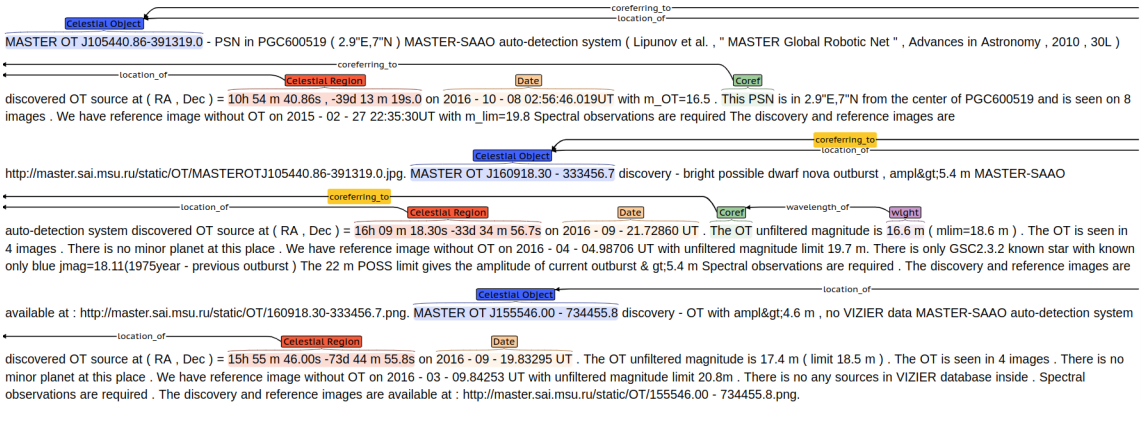
Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.



Challenges

- Example of an astronomical observation report dealing about the detection of three celestial objects.



Methodology

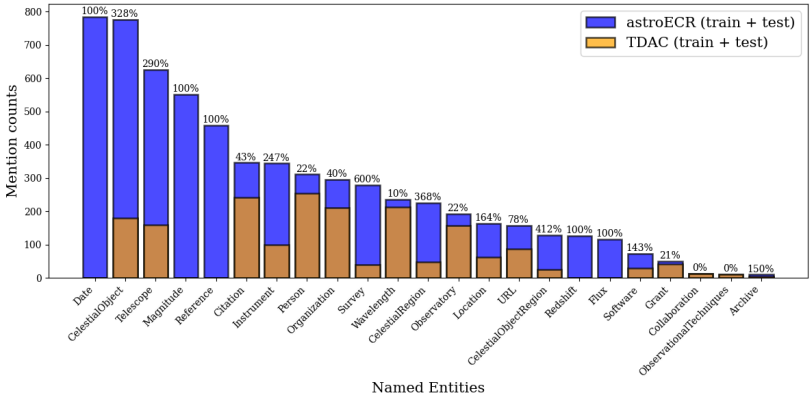
- Use TDAC corpus as a starting point:
  - **Language**: english;
  - **Nature**: observation reports (GCN Circulars, ATels and AstroNotes from the TNS);
  - **Size**: training (59 doc.), and test (16 doc.).
- Building astroECR by:
  - Expanding the TDAC corpus size: passing from 75 to 300 documents;
  - Defining 5 additional named entity categories;
  - Including coreference and semantic relation annotations;
  - Linking/normalizing celestial object names.

Content

Defining 5 Additional Named Entity Categories

- **Non-numerical-type categories:**
  - Date: *We report the discovery of a probable nova in M31 on a co-added 990-s R-band CCD frame taken under poor conditions on* **2019 Mar. 12.791 UT**$_{\text{[Date]}}$ *with the 0.65-m telescope at Ondrejov.*
  - Reference: *In comparison to the optical region (ref: the SALT spectrum in* **ATel #3289**$_{\text{[Reference]}}$*), few strong NI lines are expected in the JHK bands.*
- **Equation and numerical-type categories**:
  - Magnitude: *As reported to CBAT, this nearby-M31 object was discovered by Koichi Itagaki at* **16.5 mag.**$_{\text{[Magnitude]}}$
  - Flux: *The flux values ranged from* **1.01 +/-0.06 E+11 cgs**$_{\text{[Flux]}}$ *to* **1.71 +/- 0.04 E+11 cgs.**$_{\text{[Flux]}}$
  - Redshift: *The host KUG 0180+227 is an E+A galaxy at* **z=0.022.**$_{\text{[Redshift]}}$

Named Entities Distribution: TDAC vs astroECR



Parameters	TDAC		astroECR	
	Train	Test	Train	Test
# documents	59	16	210	90
# tokens	15374	3638	43481	10578
# ann. tokens	4338	1014	17392	3173

Source Data	TDAC	astroECR
ATels	25	175
GCN	25	100
AstroNotes	25	25

Scope of Coreferences Annotation

- **In-scope coreferences**:

- *We discovered* **PS19did** *on MJD 58666.31 = 2019-07-02.31, at w=19.9 +/- 0.1 [...]* **The new transient source** *is in the galaxy* **UGC 11003** *[...] Adopting* **the host galaxy** *redshift z=0.03566 (NED) yields an expansion velocity [...] Followup observations of* **this intrinsically faint transient** *are encouraged.*
- *We report on the discovery and follow-up of a very bright and highly magnified microlensing event* **Gaia19bld**. *[...]* **It** *has been detected and announced by the Gaia Science Alerts program.*
- *We report on the NIR brightening of the intermediate redshift quasar* **PKS0735+17** *(z=0.424), also known as* **CGRaBSJ04738+1742**.

- **Out-of-scope coreferences**:

- *Analysis of* **the data** *is ongoing. We remind the community that all* **Swift data** *are public, and encourage* **their** *use.*
- **The observations** *continued until 2019-04-26 20:15 UT, when* **they** *were aborted to begin followup of.*
- *The estimated AB magnitude is* **17.6**. **This magnitude** *is not corrected for the host galaxy contribution.*

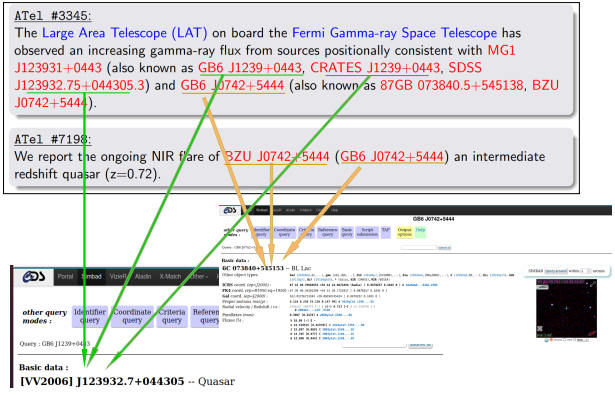
Coreferences Annotation Statistics

Parameters	**astroECR**	
	Train	**Test**
# coref. ment.	412	101
# coref. clust.	257	65
avg. clust. len.	3.5 (+/- 2.26)	3.4 (+/- 1.61)

Table: Number of coreferent mentions, number of clusters including singleton clusters, and the average cluster length (with standard deviation).

Linking Celestial Object Names in the astroECR Corpus

- Diverse naming conventions and cataloging systems in astronomy;
- We used three astronomical catalogues: SIMBAD, NED and the Transient Name Server (TNS) to link/normalize celestial object names.



Semantic Relations Annotations & Statistics

- To mainly address a within-sentence relationship detection task, we decided to link the properties described in the text to the closest mention of the celestial object's coreference in the text;

Parameters	**astroECR**	
	Train	**Test**
# within-sent rel.	490	143
# cross-sent rel.	154	26
overall ann. sem. rel.	644	169

Table: Number of annotated semantic relations in the train and test sets of our astroECR corpus with details by type of relations (within and cross-sentences).

Inter Annotator Agreement

Task	Annotators	Exact-match			Inexact-match		
		P	R	F1	P	R	F1
Named Entities	Astro vs. NLP	0,65	0,59	0,62	0,84	0,92	0.88
	Astro vs. consensus	0,83	0,86	0,84	0,93	0,96	0,94
	NLP vs. consensus	0,73	0,69	0,71	0,94	0,89	0,91
Coreferences	Astro vs. NLP	0,77	0,88	0,82	0,78	0,89	0,83
	Astro vs. consensus	0,97	1,00	0,98	0,97	1,00	0,98
	NLP vs. consensus	0,74	0,89	0,81	0,75	0,90	0,82

Table: Inter-annotator agreement for the annotation of named entities and mentions of coreference between the two annotators, and comparison with the consensus. The astrophysicist annotator is referred to as "*Astro*", and the NLP expert is referred to as "*NLP*". The metrics used are Precision (P), Recall (R), and F-measure (F1). Two evaluation modes: exact and inexact (no penalty for annotation boundaries).

Content

Named Entity Recognition Experiments & Results

Model	Precision	Recall	F1
SciBERT	0.85	0.75	0.79
astroBERT	0.83	0.81	0.82

Table: Performance of a SciBERT-based and an astroBERT-based NER system fine-tuned and evaluated on our corpus. Metrics used are Precision, Recall, and F1-score.

Category	**TDAC$_{\text{train}}$**				**100% astroECR$_{\text{train}}$**				$\Delta F1$ (%)
	N	P	R	F1	*N*	P	R	F1	
`CelestialObject`	130	0,88	0,94	0,90	519	0,94	1,0	0,97	+ 7,7
`CelestialRegion`	20	0,31	0,23	0,26	149	0,64	1,0	0,78	+ 200
`Observatory`	60	0,54	0,58	0,64	101	0,80	0,67	0,72	+ 12,49
`Database`	36	0,75	0,81	0,78	79	0,77	0,90	0,83	+ 6,4

Table: Comparison of gains obtained per class on the test set TDAC$_{\text{test}}$ depending on the training corpus used. The metrics used are Precision (P), Recall (R), and F-measure (F1). *N* corresponds to the number of mentions of entities of the class in the training corpus.

Coreference Resolution Experiments & Results

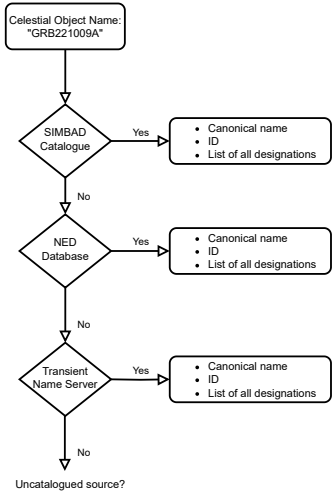
Model	CoNLL		
	Precision	Recall	F1
F-coref[1]	0.09 (0)	0.26 (0)	0.13 (0)
astroFastCoref	0.67 (0.01)	0.44 (0.01)	0.53 (0.01)

Table: Mean precision, recall, and F1-score (with standard deviation) of the F-coref system evaluated on the test set of our corpus with and without fine-tuning. Each experiment has been run five times (on 50 epochs when fine-tuning) with different random seeds.

[1] https://github.com/shon-otmazgin/fastcoref/tree/main

Entity Linking Experiments & Results

- Proposed architecture and its performances for linking/normalizing celestial object names.



Catalogue	Accuracy
SIMBAD	60.39
SIMBAD + NED	71.28
SIMBAD + NED + TNS	80.19

Relation Extraction Experiments & Results

- **Generating Negative Examples for Training:**

Sentences	Label
The median magnitude of **@@FO Aqr$_{\text{[CelestialObject]}}$**$$ *in 498 ASAS-SN observations from 2012-2015 was* **@@V=13.54$_{\text{[Magnitude]}}$**$$.	1
A spectrum was obtained using the SPRAT spectrograph on the **@@Liverpool Telescope$_{\text{[Telescope]}}$**$$. *Classification indicates it is a type Ia supernova with estimated redshift* **@@z=0.078$_{\text{[Redshift]}}$**$$	0

Table: Examples of positive (denoted by the label 1) and negative (label 0) relations. For training and inferen processes, entities in sentences are marked with the symbols @@ and $$.

- **Performance of a biLSTM binary relation detection system:**

Label	Precision	Recall	F1
1	0.77	0.80	0.79

Content

Building an Information Extraction Pipeline

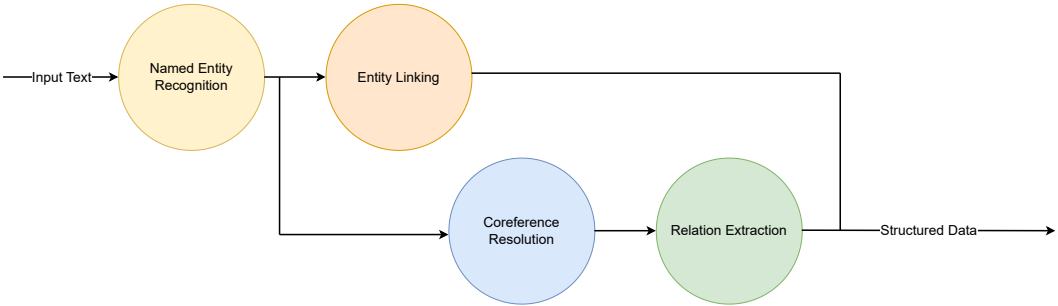
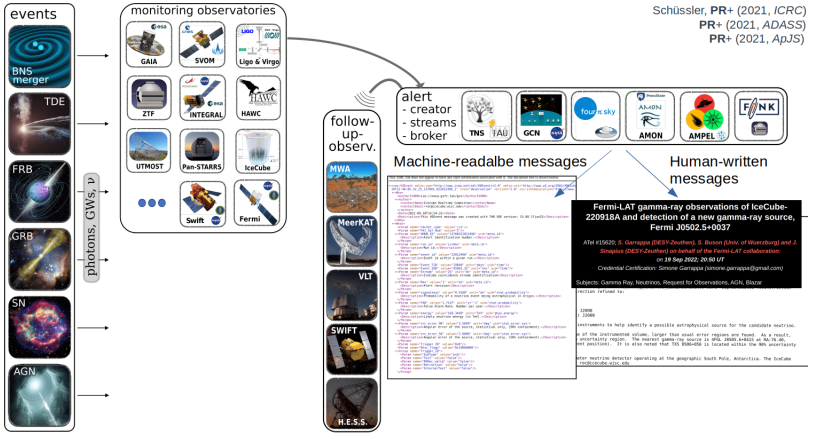


Figure: Our information extraction pipeline is composed of 4 different modules.

Deployement in Astro-COLIBRI for Real-Time Analysis of Observation Reports



Schüssler, **PR+** (2021, *ICRC*)
PR+ (2021, *ADASS*)
PR+ (2021, *ApJS*)

Deployement in Astro-COLIBRI for Real-Time Analysis

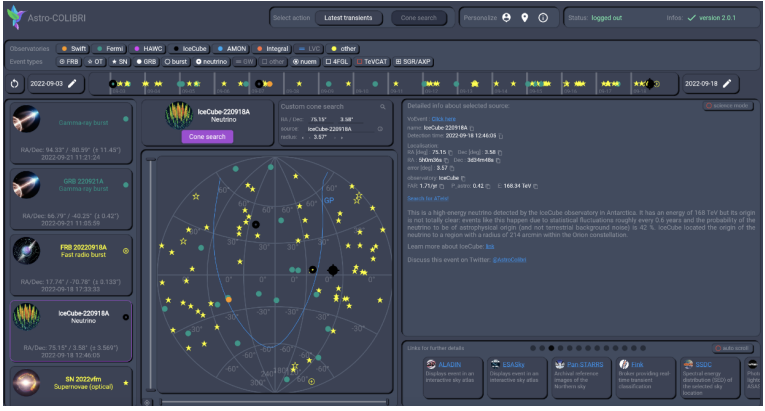


Figure: Astro-COLIBRI's web interface[2]

[2] https://astro-colibri.com/

Astro-COLIBRI: Visualize the latest transients

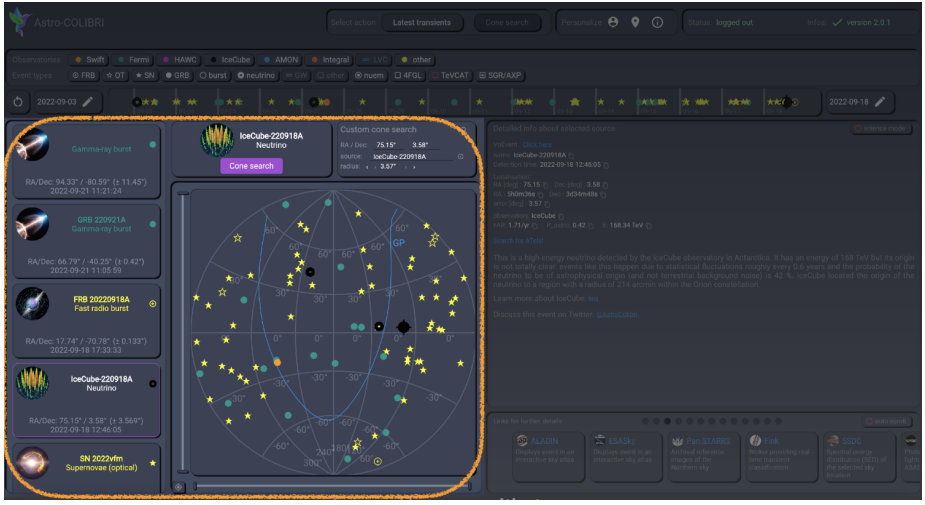


Figure: Events Map

Astro-COLIBRI: Summary for Each Transient

Figure: Information synthesis

Astro-COLIBRI: Scientific Mode for detailed information

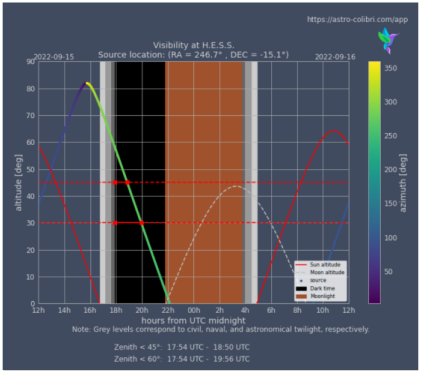


Figure: Switching to scientific mode provides access to additional information.

- The visibility plot to determine if a source is observable during the next 24 hours;
- Several observatories are implemented. Can be also computed according to our own location.

Conclusion & Additional Information

- **Web Interface**: https://astro-colibri.com
- **API** (with documentation): www.astro-colibri.science
- **Astro-COLIBRI papers**:
  https://iopscience.iop.org/article/10.3847/1538-4365/ac1517
  https://www.mdpi.com/2075-4434/11/1/22
- **Astro-COLIBRI is available for free on Android + iOS**: install it and receive real-time alert notifications!

- **contact/feedback**: astro.colibri@gmail.com

Bibliography I

Alkan et al, 2022b

TDAC, the First Time-Domain Astrophysics Corpus: Analysis and First Experiments on Named Entity Recognition.
Proceedings of the first Workshop on Information Extraction from Scientific Publications (AACL-IJCNLP 2022)

Alkan et al, 2022a

A Majority Voting Strategy of a SciBERT-based Ensemble Models for Detecting Entities in the Astrophysics Literature (Shared Task).
Proceedings of the first Workshop on Information Extraction from Scientific Publications (AACL-IJCNLP 2022)

Alkan et al, 2023

Etude de méthodes d'augmentation de données pour la reconnaissance d'entités nommées en astrophysique (Investigating Data Augmentation Methods for Astrophysical Named Entity Recognition).
Actes de la 30e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale, Paris, France, 2023.

Grezes et al, 2021

Building astroBERT, a language model for Astronomy & Astrophysics
ADASS 2021.

Grezes et al, 2022

Overview of the First Shared Task on Detecting Entities in the Astrophysics Literature (DEAL)
Proceedings of the 1st Workshop on Information Extraction from Scientific Publications.

Reichherzer et al, 2021 *ApJS* 256 5

Astro-COLIBRI—The COincidence LIBrary for Real-time Inquiry for Multimessenger Astrophysics.
The Astrophysical Journal Supplement Series, volume 256, page 5, Aug 2021