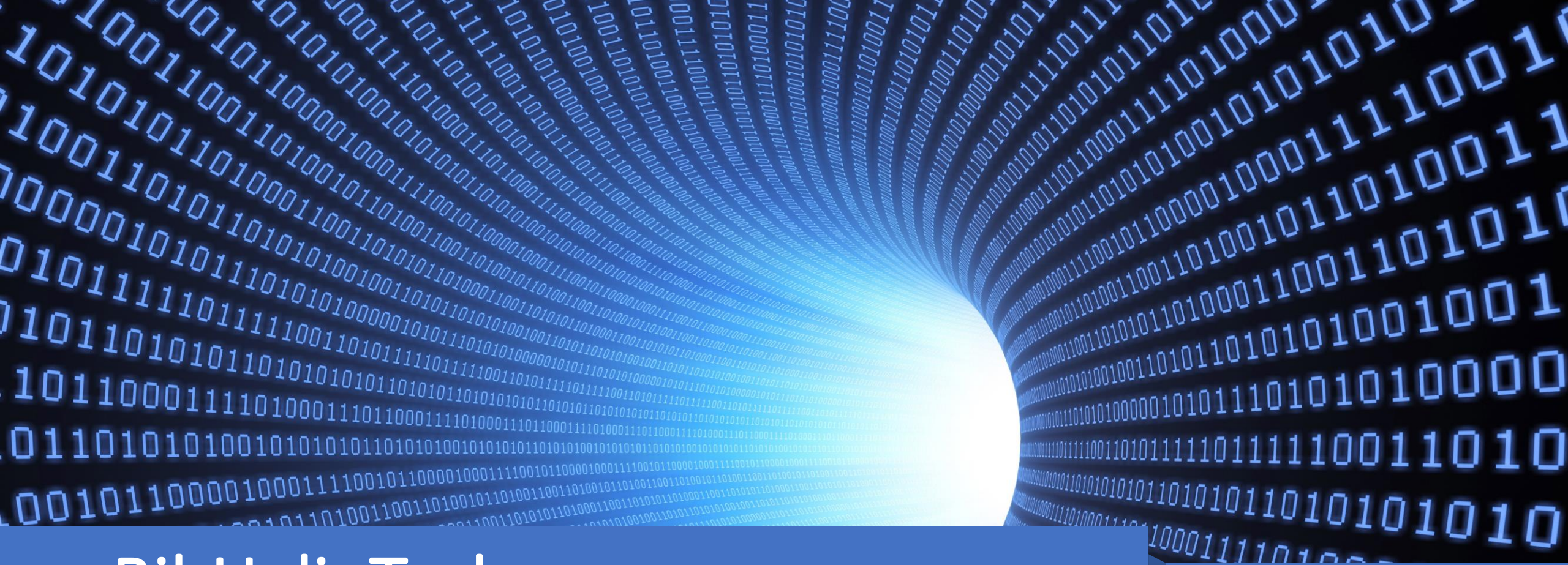


BibHelioTech

Recognition of named and temporal entities

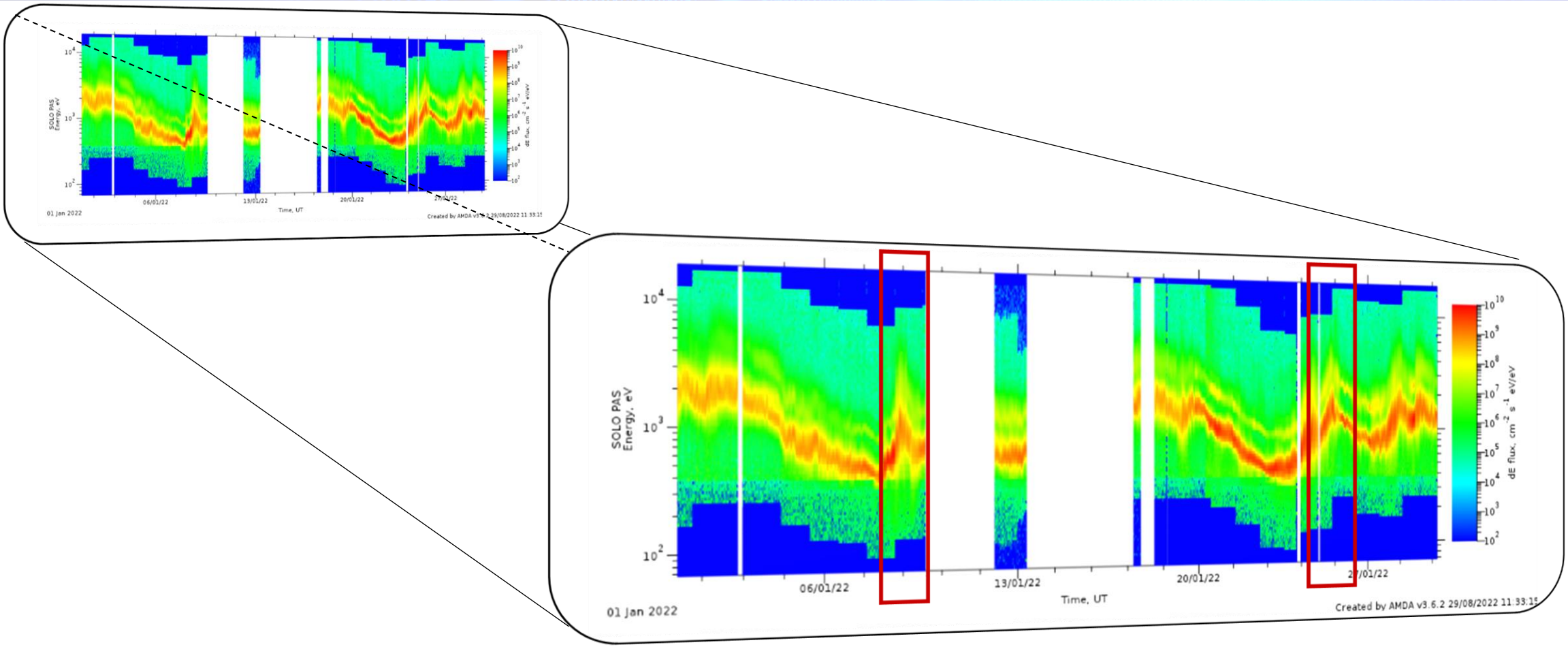
in heliophysics publications

for AMDA integration

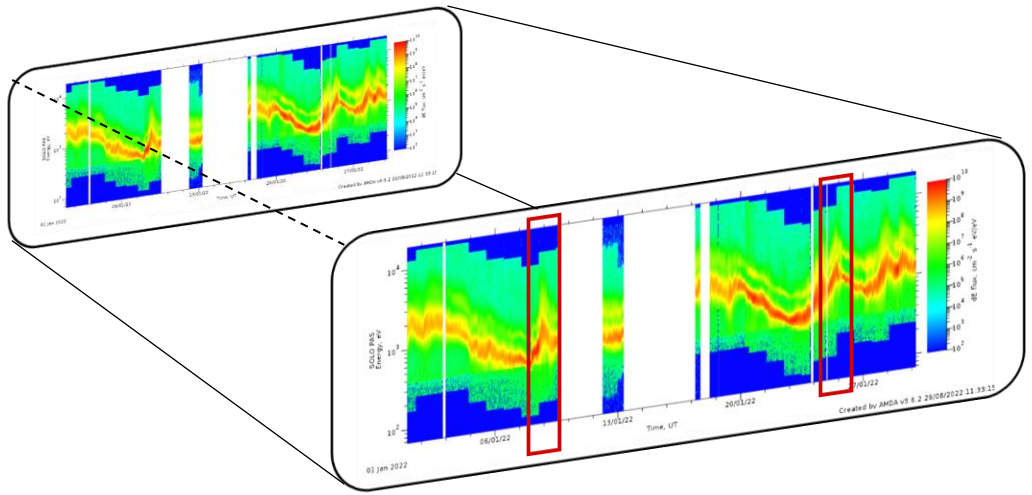
Supervised by Vincent Génot

Developed by Axel Dablanc

Needs


SOLO PAS
Energy, eV
10^4
10^3
10^2
06/01/22
13/01/22
20/01/22
27/01/22
Time, UT
01 Jan 2022
Created by AMDA v3.6.2 29/08/2022 11:33:15
dE flux, cm^{-2} s^{-1} eV/eV
10^{10}
10^9
10^8
10^7
10^6
10^5
10^4
10^3
10^2

Challenges



– Analyser un corpus d'articles scientifiques –

– Identifier et extraire les intervalles de temps, noms des missions, instruments, régions et DOI –

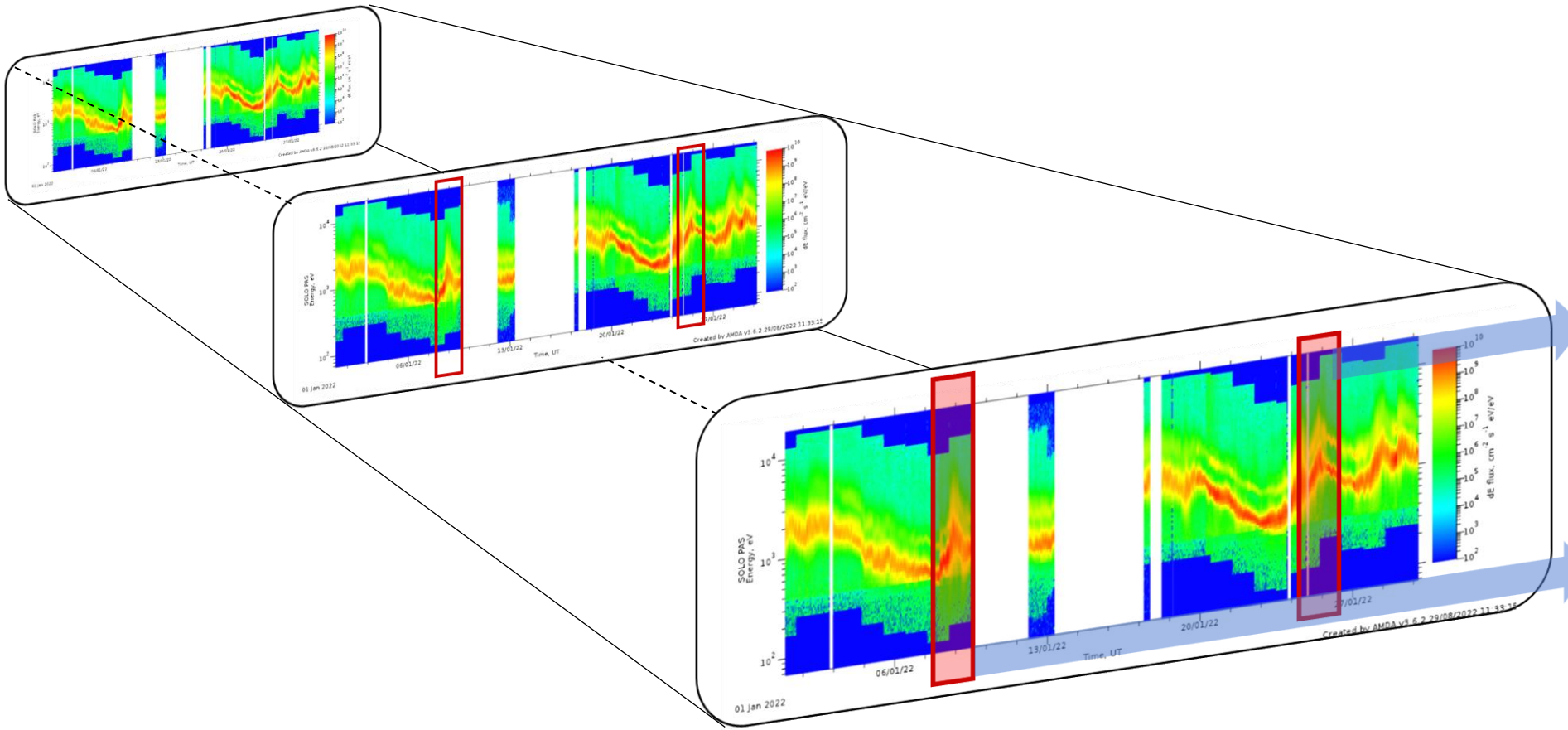
– Associer ces entités + conserver leurs cohérences –

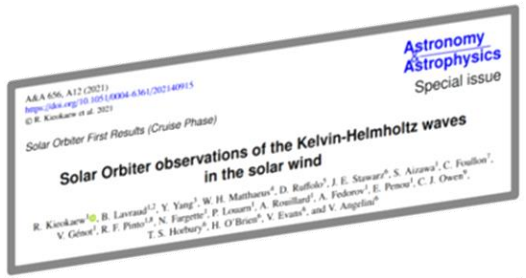
– Écrire les résultats finaux dans des fichiers utilisables par AMDA –

Figure 4. *(a) STA (top) and STB (bottom) PLASTIC energy spectrograms on 26 January to 10 March 2007. Note that on the PLASTIC energy spectra helium is clearly visible above the proton peak since the instrument switches from a larger (MC) to a smaller (SC) aperture during the high-to-low energy sweep when the count rate reaches a certain limit, which falls usually on the helium peak in the solar wind. (b) STA and STB PLASTIC energy spectrograms, SWEA electron density, and 1-minute averaged MAG magnetic field measurements during multiple bow shock crossings on 28 January to 2 February 2007. (c) STA and STB PLASTIC energy spectrograms, SWEA electron density, and 1-minute averaged MAG magnetic field measurements during multiple magnetopause crossings on 12–28 February 2007.*

50 + articles

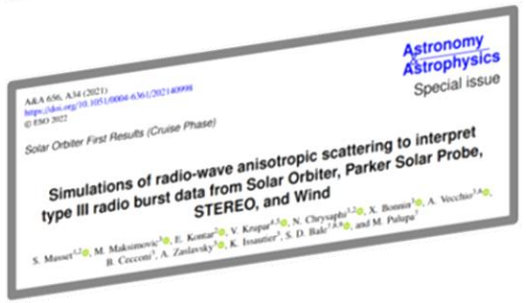
Interest




A&A 656, A12 (2021)
Astronomy & Astrophysics
Special issue
Solar Orbiter First Results (Cruise Phase)

Solar Orbiter observations of the Kelvin-Helmholtz waves in the solar wind

R. Kieokaew, B. Lavraud, Y. Yang, W. H. Matthaeus, D. Ruffolo, J. E. Stawarz, S. Aizawa, C. Foullon, V. Génot, R. F. Pinto, N. Fargette, P. Louarn, A. Rouillard, A. Fedorov, E. Penou, C. J. Owen, T. S. Horbury, H. O'Brien, V. Evans, and V. Angelini


A&A 656, A34 (2021)
Astronomy & Astrophysics
Special issue
Solar Orbiter First Results (Cruise Phase)

Simulations of radio-wave anisotropic scattering to interpret type III radio burst data from Solar Orbiter, Parker Solar Probe, STEREO, and Wind

S. Musset, M. Maksimovic, E. Kontar, V. Krupar, N. Chrysaphi, X. Bonnin, A. Vecchio, B. Cecconi, A. Zaslavsky, K. Issautier, S. D. Bale, and M. Pulupa

L’IRAP

Entity recognition: Analysis

Temporal entities

Entity recognition: Analysis

ST-A: Year DOY Month Day Hour	ST-B: Year DOY Month Day Hour
2007 65 03 06 14	2007 65 03 06 16
2007 71 03 12 12	2007 71 03 12 14
2007 84 03 25 02	2007 84 03 25 08
2007 91 04 01 02	2007 91 04 01 04
2007 98 04 08 22	2007 99 04 09 10
2007 113 04 23 02	2007 113 04 23 08
2007 117 04 27 22	2007 117 04 27 16
2007 127 05 07 14	2007 127 05 07 14
2007 138 05 18 16	2007 138 05 18 08
2007 144 05 24 16	2007 144 05 24 14
2007 155 06 04 14	2007 155 06 04 00
2007 165 06 14 18	2007 165 06 14 10
2007 173 06 22 06	2007 173 06 22 06
2007 181 06 30 08	2007 180 06 29 14
2007 185 07 04 12	2007 184 07 03 14
2007 192 07 11 14	2007 191 07 10 22
2007 196 07 15 06	2007 195 07 14 12
2007 201 07 21 04	2007 201 07 20 06
2007 208 07 27 06	2007 207 07 26 14
2007 210 07 29 12	2007 210 07 29 08
2007 219 08 07 08	2007 218 08 06 12
2007 223 08 11 04	2007 222 08 10 16
2007 238 08 26 20	2007 237 08 25 00
2007 246 09 03 06	2007 244 09 01 10
2007 250 09 07 22	2007 249 09 06 14
2007 258 09 15 06	2007 257 09 14 10
2007 264 09 21 22	2007 263 09 20 02
2007 273 09 30 02	2007 271 09 28 08
2007 277 10 04 08	2007 276 10 03 02
2007 292 10 19 08	2007 290 10 17 08
2007 299 10 26 06	2007 297 10 24 20
2007 314 11 10 16	2007 314 11 10 00
2007 318 11 14 10	2007 317 11 13 08
2007 325 11 21 04	2007 324 11 20 00
2007 330 11 26 00	2007 327 11 23 22
2007 346 12 12 16	2007 343 12 09 14
2007 353 12 19 00	2007 350 12 16 12
2007 355 12 21 00	2007 352 12 18 18

Date and time
2015-10-16 13:07:02
2015-12-08 11:20:43
2015-12-06 23:38:31
2015-10-25 11:07:46
2016-12-09 09:03:54
2016-11-09 13:39:26
2017-07-11 22:34:02
2017-06-17 20:24:07
2017-08-10 12:18:33
31 EDRs

from April 27, 2019 to May 24, 2019 and from May 23, 2019 to June 19, 2019

orbit 14767 and MAVEN orbit 1754 obtained during Day Of Year (DOY) 235 and 236 in 2015. In Figures 1a and b, the energy spectrum and the density and speed of the pris-

Day in 1984	Start, UT	Stop, UT
Sept. 1	0642:00	0718:00
Sept. 21	1313:00	1336:00
Sept. 21	1426:00	1444:00
Sept. 25	0624:35	0626:30
Sept. 25	0648:00	0654:00

on 21 January 2016 01:06:41.10–01:06:52.04 UT,

Pc 5 plasma oscillations from1030 to1400 in the early morning sector, similar to the events occurring on April 29 1996. Peak frequencies were 2.6 mHz, and 3.3 mHz during the intervals of 1030 - 1300 and 1300 - 1400, respectively. In Figure 4a are seen several sheath entries of a few minute duration during the period from 1600 to 1800. Large amplitude oscillations of more than 200 km/s in plasma flow were observed during the intervals of 1530 - 1600 and 1655 - 1740. In the latter period temporal entries of the spacecraft into the magnetosheath were clearly noted in the

2015-10-05/06:34:00.00 to 2015-10-05/06:36:20.00 UT [X, Y, Z, total] = [−9.8, 30.9, 9.5, 33.8] nT 6.0	2015-10-05/06:36:20.00 to 2015-10-05/06:37:00.00 UT [X, Y, Z, total] = [−3.5, 11.9, −15.7, 20.1] nT 14.6
Mirror-modulated current sheet	
2015-10-05/06:36:15.00 to 2015-10-05/06:36:17.00 UT [X, Y, Z, total] = [−5.7, 42.7, 23.8, 49.2] nT 1.3	2015-10-05/06:36:22.00 to 2015-10-05/06:36:24.00 UT [X, Y, Z, total] = [0.3, 9.3, −8.9, 12.8] nT 34.5

Entity recognition: Temporal

Text read

SUTime results


```
"from April 27, 2019 to May 24, 2019 and from May 23, 2019 to June 19, 2019"
[
    {
        "end": 35,
        "start": 0,
        "text": "from April 27, 2019 to May 24, 2019",
        "type": "DURATION",
        "value": {
            "begin": "2019-04-27",
            "end": "2019-05-24"
        }
    },
    {
        "end": 74,
        "start": 40,
        "text": "from May 23, 2019 to June 19, 2019",
        "type": "DURATION",
        "value": {
            "begin": "2019-05-23",
            "end": "2019-06-19"
        }
    }
]
```

Entity recognition: Temporal

SUTime results

```
"21 January 2016 01:06:41.10-01:06:52.04"
[
    {
        "end": 39,
        "start": 0,
        "text": "21 January 2016 01:06:41.10\u201301:06:52",
        "type": "DURATION",
        "value": {
            "begin": "2016-01-21T01:06:41",
            "end": "T01:06:52"
        }
    }
]
```

Text read

Entity recognition: Temporal

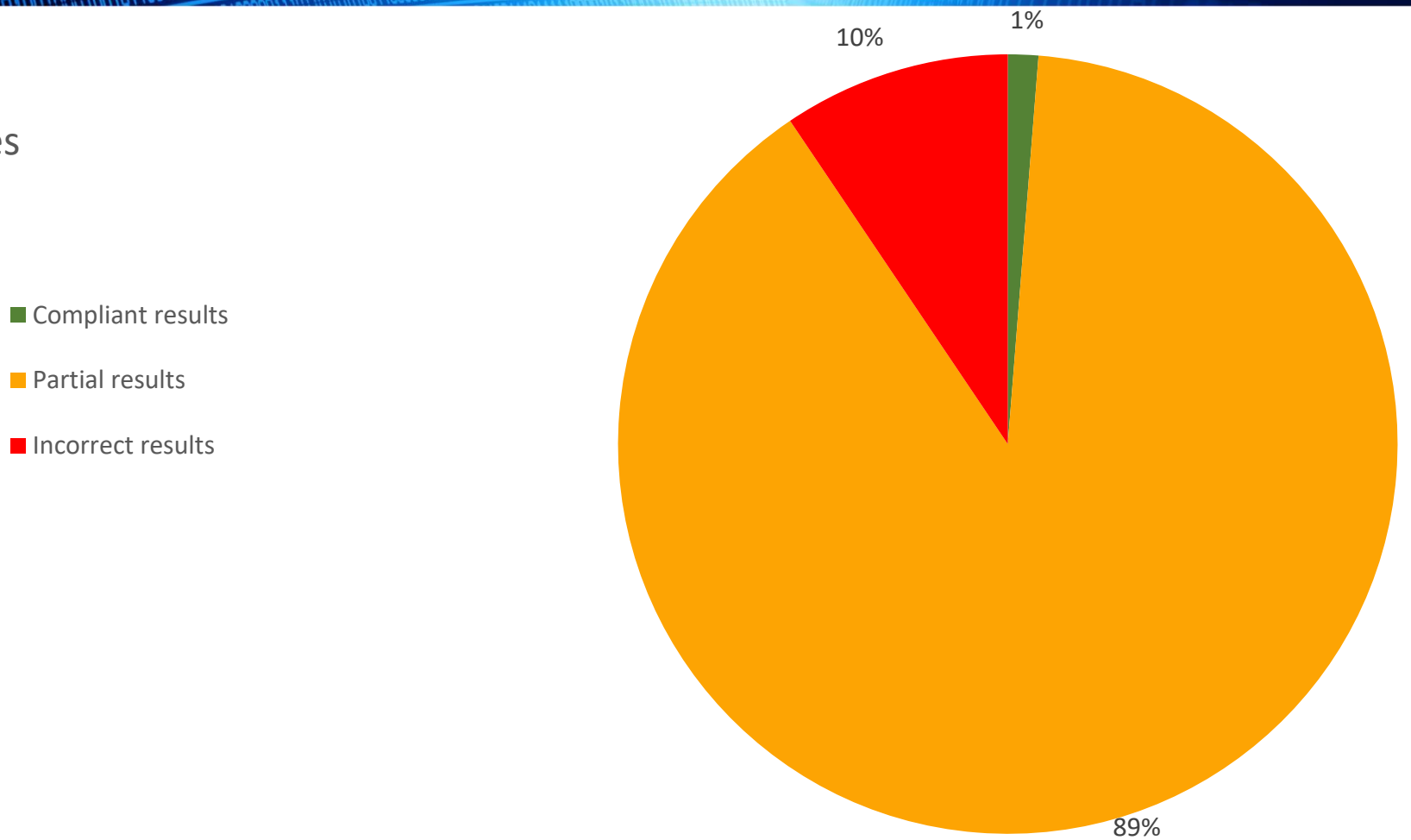
Text read

SUTime results


```
"1030 - 1300 and 1300 - 1400"
[
    {
        "end": 11,
        "start": 0,
        "text": "1030 - 1300",
        "type": "DURATION",
        "value": {
            "begin": "1030",
            "end": "1300"
        }
    },
    {
        "end": 27,
        "start": 16,
        "text": "1300 \u2013 1400",
        "type": "DURATION",
        "value": {
            "begin": "1300",
            "end": "1400"
        }
    }
]
```

Entity recognition: Temporal

SUTime results
before
Filtering, transformations and adding rules



Paper reading: Design

Filtering and Transformations

```
content = re.sub(r'References\n[\s\S]+', '', content)
content = re.sub(r'REFERENCES\n[\s\S]+', '', content)
content = re.sub('Received.*', '', content)
content = re.sub('Accepted.*', '', content)
content = re.sub('Published.*', '', content)
content = re.sub('Suggested.*', '', content)

content = re.sub("\n{1,5}"," ", content) # remove all \n

content = re.sub(r"UT ", r" UT ", content)  # replace "22:02UT" by "22:02 UT"
content = re.sub(r" UT ", r"UTC", content)  # replace "22:02 UT" by "22:02 UTC"

# HHmm to HH:mm
content = re.sub(r"([0-9]{2})([0-9]{2})(\:[0-9]{2})", r"\1:\2\3", content)
content = re.sub(r"([0-9]{2})([0-9]{2})((-|-|–)|(-|-|–) | (-|-|–)| (-|-|–) | to | and )([0-9]{2})([0-9]{2})", r"\1:\2 - \8:\9", content)
content = re.sub(r"([0-9]{2})([0-9]{2}) (UTC)", r"\1:\2\3", content)

content = re.sub(
    r"((?:Jan(?:uary)?|Feb(?:ruary)?|Mar(?:ch)?|Apr(?:il)?|May|Jun(?:e)?|Jul(?:y)?|Aug(?:ust)?|Sep(?:tember)?|Oct(?:ober)?|(Nov|Dec)(?:ember
    r"\1 \3 \7", content)  # replace "09:30-23:30" by "09:30 - 23:30"
content = re.sub(r"([0-9]{2}\:[0-9]{2})(\-|\-|\–)([0-9]{2}\:[0-9]{2})", r"\1 \2 \3", content)  # replace "09:30-23:30" by "09:30 - 23:30"
content = re.sub(r"([0-9]{2}\:[0-9]{2})(\-|\-|\–) ([0-9]{2}\:[0-9]{2})", r"\1 \2 \3", content)  # replace "09:30- 23:30" by "09:30 - 23:30"
content = re.sub(r"([0-9]{2}\:[0-9]{2}) (\-|\-|\–)([0-9]{2}\:[0-9]{2})", r"\1 \2 \3", content)  # replace "09:30 -23:30" by "09:30 - 23:30"
content = re.sub(r"([0-9]{2}\:[0-9]{2}\:[0-9]{2}\.[0-9]{1,3})", r"T\1", content)  # replace "10:28:42.950" by "T10:28:42.950"
```

Paper reading: Design

Add rules

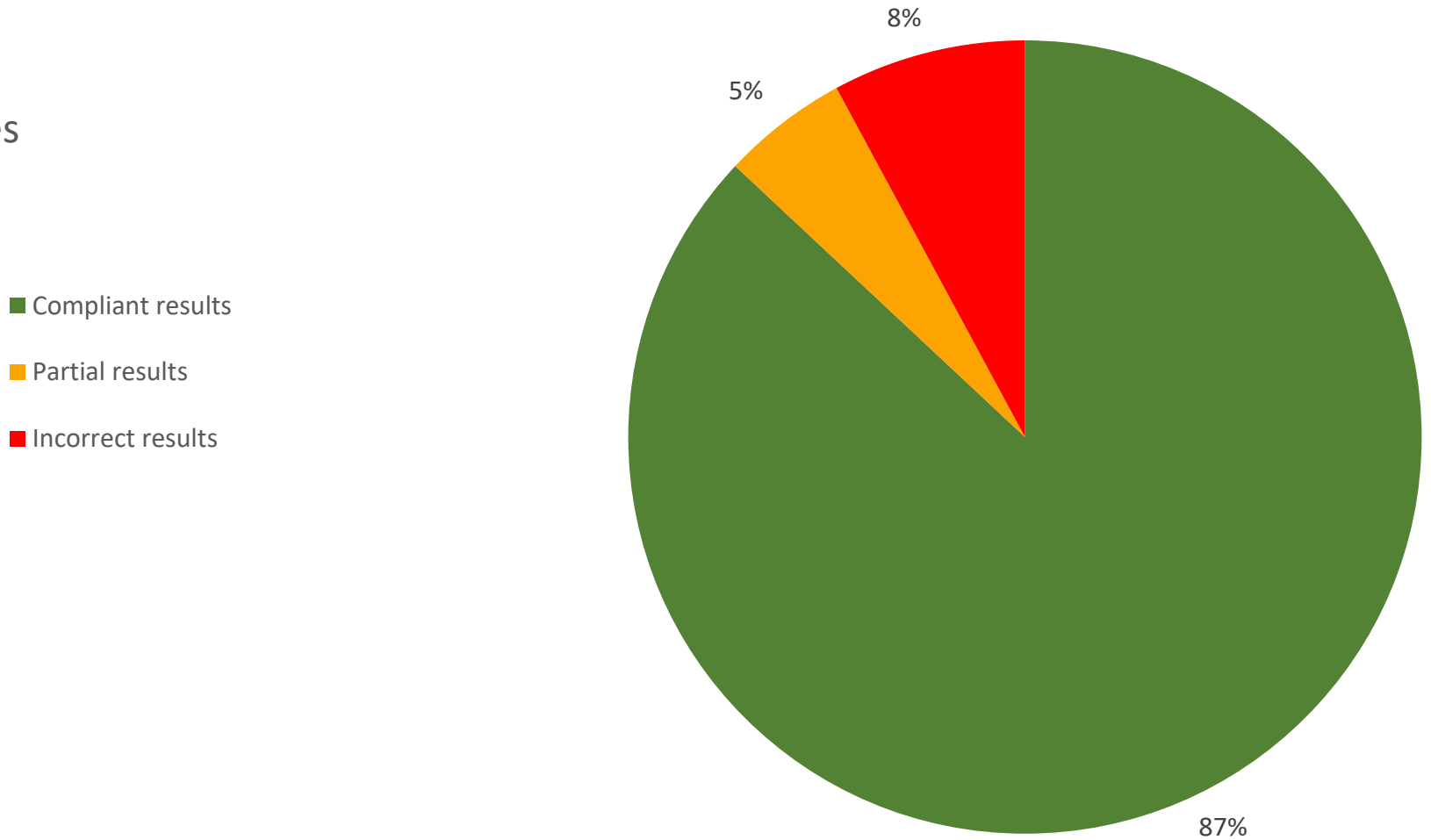
```
{ name: "temporal-composite-8:ranges",
active: options."markTimeRanges",
  pattern: ( /from/? ( [ { temporal::IS_TIMEX_TIME } | { temporal::IS_TIMEX_DATE } ] ) /to|-|–|—|until|~/ ( [ { temporal::IS_TIMEX_TIME } |   {   temporal::IS_TIMEX_DATE } ] ) ),
result: TimeRange( $1[0].temporal.value, $2[0].temporal.value ) }

  { name: "temporal-composite-8a:ranges",
active: options."markTimeRanges",
  pattern: ( /during/? ( [ { temporal::IS_TIMEX_TIME } | { temporal::IS_TIMEX_DATE } ] ) /to|-|–|until|~/ ( [ { temporal::IS_TIMEX_TIME } |   {   temporal::IS_TIMEX_DATE } ] ) ),
result: TimeRange( $1[0].temporal.value, $2[0].temporal.value ) }

{ name: "temporal-composite-8b:ranges",
active: options."markTimeRanges",
  pattern: ( /between/? ( [ { temporal::IS_TIMEX_TIME } | { temporal::IS_TIMEX_DATE } ] ) /and|–/ ( [ { temporal::IS_TIMEX_TIME } |   {   temporal::IS_TIMEX_DATE } ] ) ),
result: TimeRange( $1[0].temporal.value, $2[0].temporal.value ) }
```

Entity recognition: Temporal

SUTime results
after
Filtering, transformations and adding rules



Entity recognition: Analysis

Named entities

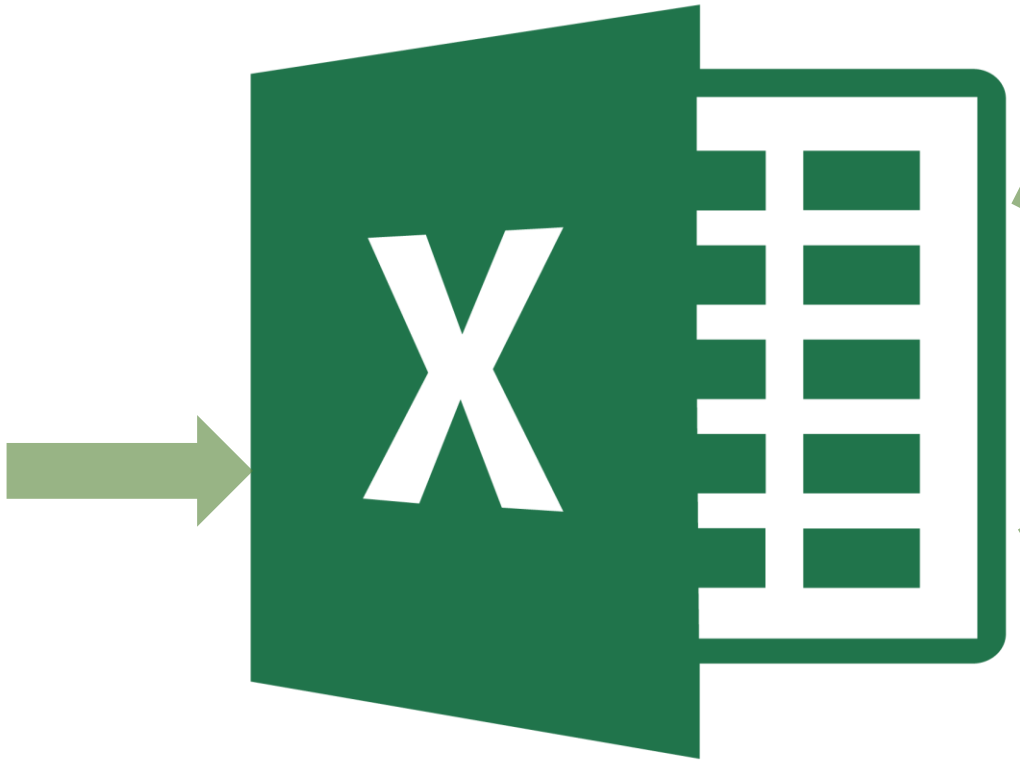
Entity recognition: Analysis

Artificial Intelligence

Text search

Entity recognition: Named

Satellites
Names and synonyms
+
Periods of existence
+
Observation regions



Instruments
Names and acronyms

Space Regions

Entity recognition: Named

NOM	SYNONYME 1	SYNONYME 2	SYNONYME 3	SYNONYME 4	SYNONYME 5
ABOVE					
ACE	Advanced Composition Explorer				
AE C	AE-C	Atmospheric Exp	AEC		
AE D	AE-D	Atmospheric Exp	AED		
AE E	AE-E	Atmospheric Exp	AEE		
AE		Atmospheric Explorer			
AIM	Aeronomy of Ice in the Mesosphere				
AMPTE CCE	CCE	AMPTE-CCE	Active Magnetos	AMPTECCE	
AMPTE IRM	IRM	AMPTE-IRM	Active Magnetos	AMPTEIRM	
AMPTE UKS	UKS	AMPTE-UKS	Active Magnetos	AMPTEUKS	
AMPTE	Active Magnetospheric Particle Tracer Explorers				
ARTEMIS	Advanced Relay and Technology Mission				
AUTUMN					
AWESOME					
Aeros A	Aeros-A	AerosA			

Excel sheet:
Satellites

Entity recognition: Named

NOM	INSTRUMENT 1	INSTRUMENT 2	INSTRUMENT 3	INSTRUMENT 4	INSTRUMENT 5	INSTRUMENT 6	INSTRUMENT 7	INSTRUMENT 8	INSTRUMENT 9	INSTRUMENT 10	INSTRUMENT 11	INSTRUMENT 12
ACE	CRIS	EPAM	{'MAG':'Magnetc	SEPICA	{'SIS':'Supratherr	{'SWEPAM':'Sola	{'SWICS':'Solar W	SWIMS	ULEIS	{'MFI':'Wind Magnetic Field Investigation'}		
AE-C	BIMS	CEP	LEE	MESA	MIMS	NACE	NATE	OSS	PES	{'RPA':'Reme Co	UVNO	VAE
AE-D	CEP	LEE	MESA	MIMS	NACE	NATE	OSS	PES	{'RPA':'Reme Co	UVNO	VAE	
AE-E	BIMS	BUV	CEP	ESUM	EUVS	MESA	NACE	NATE	OSS	PES	{'RPA':'Reme Co	VAE
AE C	BIMS	CEP	LEE	MESA	MIMS	NACE	NATE	OSS	PES	{'RPA':'Reme Co	UVNO	VAE
AE D	CEP	LEE	MESA	MIMS	NACE	NATE	OSS	PES	{'RPA':'Reme Co	UVNO	VAE	
AE E	BIMS	BUV	CEP	ESUM	EUVS	MESA	NACE	NATE	OSS	PES	{'RPA':'Reme Co	VAE

Excel sheet:
Instruments

Entity recognition: Named

REGIONS										
Asteroid										
Comet										
Earth	{'Magnetosheath':'Magi	{'Magnetosphere':{'Ma	{'Moon':'Moon'}	{'NearSurface':{'Atmosp	{'Surface':'Surface'}					
Interstellar										
Jupiter	{'Callisto':'Callisto'}	{'Europa':'Europa'}	{'Ganymede':'Ganymed	{'Io':'Io'}	{'Magnetosphere':{'Magnetotail':'Magnetotail','Main':'Main','Plasmasphere':'Plasmasphere','Polar':'Polar','RadiationBelt':'RadiationBelt','RingCurrent':'RingCurrent'}}					
Mars	{'Deimos':'Deimos'}	{'Magnetosphere':{'Ma	{'Phobos':'Phobos'}							
Mercury	{'Magnetosphere':{'Magnetotail':'Magnetotail','Main':'Main','Plasmasphere':'Plasmasphere','Polar':'Polar','RadiationBelt':'RadiationBelt','RingCurrent':'RingCurrent'}}									
Neptune	{'Magnetosphere':{'Ma	{'Proteus':'Proteus'}	{'Triton':'Triton'}							
Saturn	{'Dione':'Dione'}	{'Enceladus':'Enceladus	{'Iapetus':'Iapetus'}	{'Magnetosphere':{'Ma	{'Mimas':'Mimas'}	{'Rhea':'Rhea'}	{'Tethys':'Tethys'}	{'Titan':'Titan'}		
Sun	{'Chromosphere':'Chror	{'Corona':'Corona'}	{'Interior':'Interior'}	{'Photosphere':'Photos	{'TransitionRegion':'TransitionRegion'}					
Uranus	{'Ariel':'Ariel'}	{'Magnetosphere':{'Ma	{'Miranda':'Miranda'}	{'Oberon':'Oberon'}	{'Puck':'Puck'}	{'Titania':'Titania'}	{'Umbriel':'Umbriel'}			
Venus	{'Magnetosphere':{'Magnetotail':'Magnetotail','Main':'Main','Plasmasphere':'Plasmasphere','Polar':'Polar','RadiationBelt':'RadiationBelt','RingCurrent':'RingCurrent'}}									
Pluto										
Heliosphere	{'Heliosheath':'Heliosh	{'Inner':'Inner'}	{'NearEarth':'NearEarth	{'Outer':'Outer'}	{'Remote1AU':'Remote1AU'}					

Excel sheet:
Spatial regions

Entity recognition: Named

A	B
NAME	**HPDE/AMDA_REG**
ABOVE	['Earth']
ACE	['Heliosphere.NearEarth','Heliosphere.Inner']
AE	['Earth.NearSurface']
AE-C	['Earth.NearSurface']
AE-D	['Earth.NearSurface']
AE-E	['Earth.NearSurface']
Aeros-A	['Earth.NearSurface']
AIM	['Earth.NearSurface']
Akebono	['Earth.Magnetosphere']
Alouette1	['Earth.Magnetosphere']
Alouette2	['Earth.Magnetosphere']
AMPTE	['Earth.Magnetosheath','Earth.Magnetosphere','Earth.Magnetosphere.Magnetotail','Heliosphere.NearEarth']

Excel sheet:
Spatial regions by satellite

Entity recognition: Named

NOM	StartDate	StopDate
ABOVE	2014-06-25	
ACE	1997-08-25	
AE C	1973-12-16	
AE D	1975-11-20	
AE E	1975-11-20	
AE		
AIM	2007-04-25	
AMPTE CCE	1984-08-16	1989-01-11
AMPTE IRM	1984-08-16	1986-08-14
AMPTE UKS	1984-08-25	1985-06-15
AMPTE		

Excel sheet:
Satellite mission period

Reviews

- Regex ↑ -
- APIs ↑ -
- Data formatting and structuring ↑ -
- Project management ↑ -
- Scanning + -
- Entity recognition + -

Reviews

```
# Name: 103389fspas2021718024_bibheliotech_V1;
# Creation Date: 2022-07-20T11:39:10.934694;
# Description: Catalogue of events resulting from the HelioNER code (Dablanc & Génot, "https://github.com/ADablanc/BibHelioTech.git") on the paper "https://doi.org/10.3389/fspas.2021.718024". The two
first columns are the start/stop times of the event. the third column is the DOI of the paper, the fourth column is the mission that observed the event with the list of instruments (1 or more) listed in
the fifth column. The sixth column is the most probable region of space where the observation took place (SPASE ObservedRegions term);
# Parameter 1: id:column1; name:DOI; size:1; type:char;
# Parameter 2: id:column2; name:SATS; size:1; type:char;
# Parameter 3: id:column3; name:INSTS; size:1; type:char;
# Parameter 4: id:column4; name:REGS; size:1; type:char;
# Parameter 5: id:column5; name:D; size:1; type:int;
# Parameter 6: id:column6; name:R; size:1; type:int;
# Parameter 7: id:column7; name:SO; size:1; type:int;
# Parameter 8: id:column8; name:occur_sat; size:1; type:int;
# Parameter 9: id:column9; name:nb_durations; size:1; type:int;
# Parameter 10: id:column10; name:conf; size:1; type:float;
2021-01-01T00:00:00.000 2021-06-30T23:59:59.000 https://doi.org/10.3389/fspas.2021.718024 "BepiColombo" "MAG,MEA1,ENA" "Mercury" 5400 1 67 174 6 0.0023428002710706684
2021-01-01T00:00:00.000 2021-06-30T23:59:59.000 https://doi.org/10.3389/fspas.2021.718024 "BepiColombo" "MAG,MEA1,ENA" "Mercury" 4163 1 67 174 6 0.0018061254682346655
2021-01-01T00:00:00.000 2021-06-30T23:59:59.000 https://doi.org/10.3389/fspas.2021.718024 "BepiColombo" "MAG,MEA1,ENA" "Mercury" 3406 1 67 174 6 0.0014776995783827218
```

Temps début | Temps fin | DOI | Nom satellite | Instruments | Région | Mesures diverses

Extract from entity recognition results

Reviews

	Start Time	Stop Time	Duration (Min)	DOI	SATS	INSTS	REGS	D	R	SO	occur...	nb_d...	conf
1	1993-06-01T00:00:00	1994-02-28T23:59:59	393179.983333	https...	"Voya...	""	"Heli...	37568	1	24	268	92	0.015...
2	1993-06-01T00:00:00	1994-02-28T23:59:59	393179.983333	https...	"Voya...	"MAG...	"Heli...	37568	1	19	268	92	0.019...
3	2014-05-07T16:30:00	2014-05-07T16:30:59	0.983333	https...	"Mars...	""	"Mars"	22948	1	12	268	92	0.018...
4	2014-05-07T16:30:00	2014-05-07T16:30:59	0.983333	https...	"Mars...	""	"Mars"	23773	1	12	268	92	0.019...
5	2014-05-07T16:30:00	2014-05-07T16:30:59	0.983333	https...	"Mars...	""	"Eart...	22964	1	12	268	92	0.018...
6	2014-05-07T17:30:00	2014-05-07T17:30:59	0.983333	https...	"Cass...	"MAG...	"Ven...	23342	1	14	268	92	0.016...
7	2014-05-07T17:30:00	2014-05-07T17:30:59	0.983333	https...	"Neut...	"Bart...	"Eart...	23709	1	3	268	92	0.076...
8	2014-05-07T17:30:00	2014-05-07T17:30:59	0.983333	https...	"Rose...	"MAG...	"Com...	23726	1	32	268	92	0.007...
9	2014-10-01T20:48:00	2014-10-01T20:48:59	0.983333	https...	"STE...	""	"Heli...	1200	1	5	268	92	0.002...
10	2014-10-01T21:39:00	2014-10-01T21:39:59	0.983333	https...	"New...	"SWA...	"Jupit...	1689	1	46	268	92	0.000...
11	2014-10-01T21:39:00	2014-10-01T21:39:59	0.983333	https...	"STE...	"EUV...	"Heli...	1658	1	36	268	92	0.000...
12	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Pion...	""	"Heli...	21990	1	3	268	92	0.070...
13	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Ulys...	"LET"	"Heli...	21064	1	3	268	92	0.067...
14	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Ulys...	"LET"	"Heli...	21284	1	3	268	92	0.068...
15	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Voya...	""	"Heli...	19337	1	24	268	92	0.007...
16	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Voya...	"MAG...	"Heli...	19032	1	19	268	92	0.009...
17	2014-10-07T14:30:00	2014-10-07T14:30:59	0.983333	https...	"Voya...	"MAG...	"Heli...	19337	1	19	268	92	0.009...
18	2014-10-07T16:30:00	2014-10-07T16:30:59	0.983333	https...	"Pion...	"MAG...	"Heli...	21847	1	1	268	92	0.210...
19	2014-10-14T00:00:00	2015-04-14T23:59:59	263519.983333	https...	"MAV...	"MAG...	"Mars"	9779	1	18	268	92	0.005...
20	2014-10-14T00:00:00	2015-04-14T23:59:59	263519.983333	https...	"MAV...	"MAG...	"Mars"	9886	1	18	268	92	0.005...
21	2014-10-14T00:00:00	2015-04-14T23:59:59	263519.983333	https...	"MAV...	"MAG...	"Mars"	9929	1	18	268	92	0.005...
22	2014-10-14T00:00:00	2015-04-14T23:59:59	263519.983333	https...	"MAV...	"MAG...	"Mars"	9944	1	18	268	92	0.005...

Catalog extract readable by AMDA

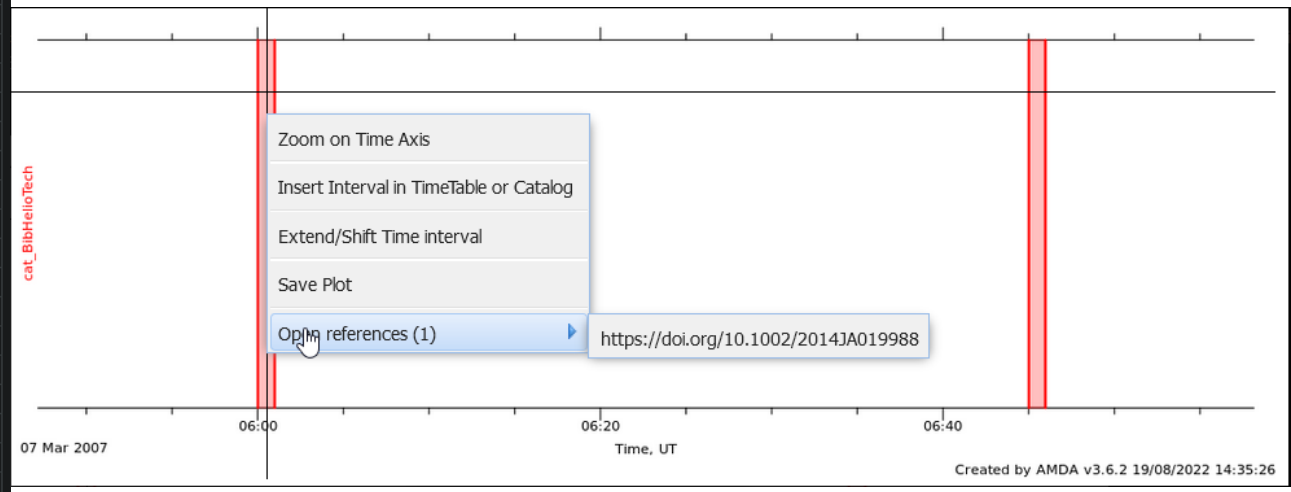
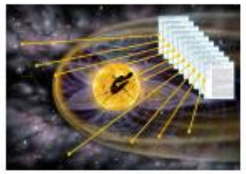


Illustration interface DOI, AMDA

PNST Physics Symposium May 2022

Vincent Génot, Arel Dablanc, Guillaume Cabanac, INSU/IRAP et IRIT (Toulouse)
Camille de Salabert, Sabine Barreaux, Pascal Cuzac, Inist-CNRS (Nancy)
Nicolas Dufourg, CNES/DOA (Toulouse)
Nicolas Aunai, INSIS/INSU/LPP (Palaiseau)
Williams Exbrayat, INSU/Observatoire Midi-Pyrénées/UAR831 (Toulouse)

BibHelioTech

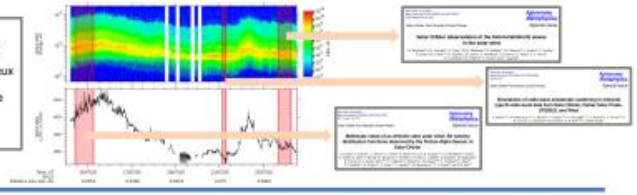


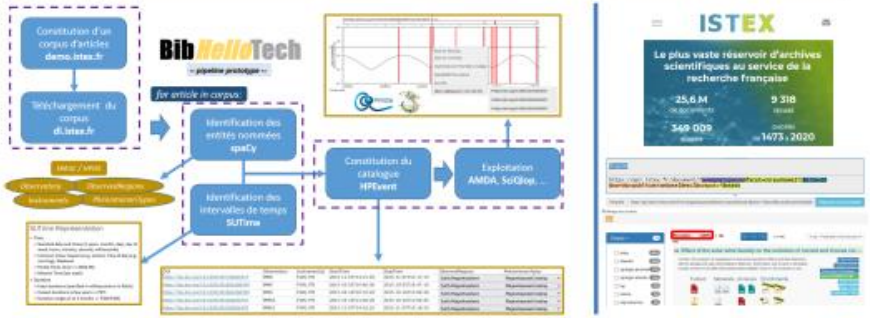
A partir d'un corpus d'articles scientifiques du domaine héliophysique utilisant des données de missions spatiales, nous réaliserons une détection textuelle automatisée sur les événements observés, les satellites/instruments utilisés, les régions spatiales et les processus physiques concernés, afin de relier ces entités avec les publications dont elles sont extraites, dans des catalogues exploitables par les outils d'analyse de données de la discipline. Ce lien fort et systématisé entre données et publications, inexistant à ce jour, augmentera l'expérience d'analyse de données en immergeant le chercheur dans le contexte bibliographique de son cas d'étude, améliorera significativement la reproductibilité des résultats publiés, et facilitera la réutilisation de ces catalogues dans de nouvelles études statistiques et comparatives.

publi data

Cas d'utilisation

- Visualisation de données Solar Orbiter PAS dans **AMDA** en Juillet 2020. Les 3 intervalles en rouge correspondent à ceux étudiés dans les 3 articles à droite.
- Le but de **BibHelioTech** est de produire des catalogues, disponibles dans les outils, liant publications et intervalles d'étude par les missions/instruments.

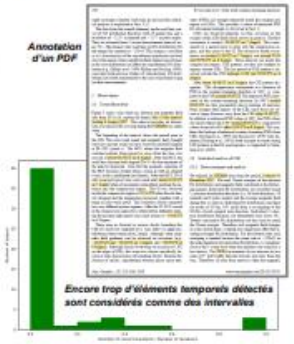




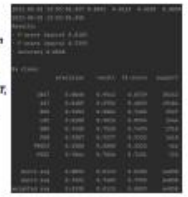
Etapes / avancement du projet (1.5 mois du stage d'A. Dablanc)

- Annotation de +50 articles en se focalisant sur **intervalles de temps, missions et instruments** (tels que dispo sur le registre SPASE https://hpde.io/SMWG/index.html)
  - Récupération automatique des annotations (librairie **PyMuPDF**)
- **Océrisation** = PDF → image → TXT :
  - **TESSERACT** (librairie python qui convertit le PDF en texte brut)
    - Lourd, mais toutes les infos sont conservées
  - **GROBID** fournit un fichier XML structuré
    - Rate beaucoup de données, dont les images et légendes
- Reconnaissance automatique des **DOI**, 2 approches complémentaires efficaces :
  - GROBID à partir de la balise DOI
  - Si DOI absent de GROBID : à partir du titre de l'article et des API de NASA / ADS
- **SUTime** : reconnaissance d'entités temporelles, utilisé pour les intervalles de temps
  - Performant (balise « duration »), mais pas encore optimisé
  - Métriques obtenues à partir d'une comparaison avec les intervalles annotés
- **NER (Named Entity Recognition)**, 2 approches :
  - utilisation de **FLAIR** (librairie python)
    - 1er essai d'entrainement d'un modèle avec de nouvelles entités ***INST, SAT, REG, PROCS***
    - Performance (entrainement sans contexte) : *voir métriques* →
  - Comparaison textuelle (plus légère en temps de traitement/CPU)
- **TODO** : association entre les intervalles et missions/instruments détectés
  - *C'est une des difficultés identifiées !*
  - Nous testerons une approche statistique en fonction de l'occurrence entre éléments et de leur proximité dans l'article

Annotation d'un PDF



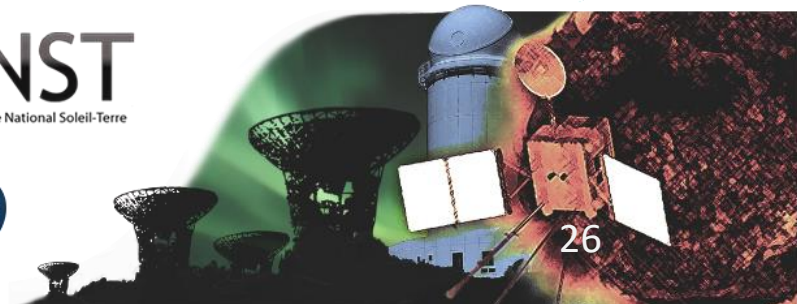
Résultat de l'utilisation de la librairie FLAIR pour l'entrainement d'un modèle avec les nouvelles classes INST, SAT, REG, PROCS issues du modèle de données SPASE



cnes cnrs Université Toulouse III Paul Sabatier irap OMP IRIT LATMOS CDPP Inist

PNST Programme National Soleil-Terre

cnrs



BibHelioTech's future

- Recruitment experienced developer -
- Handle a large database (ISTEX) -

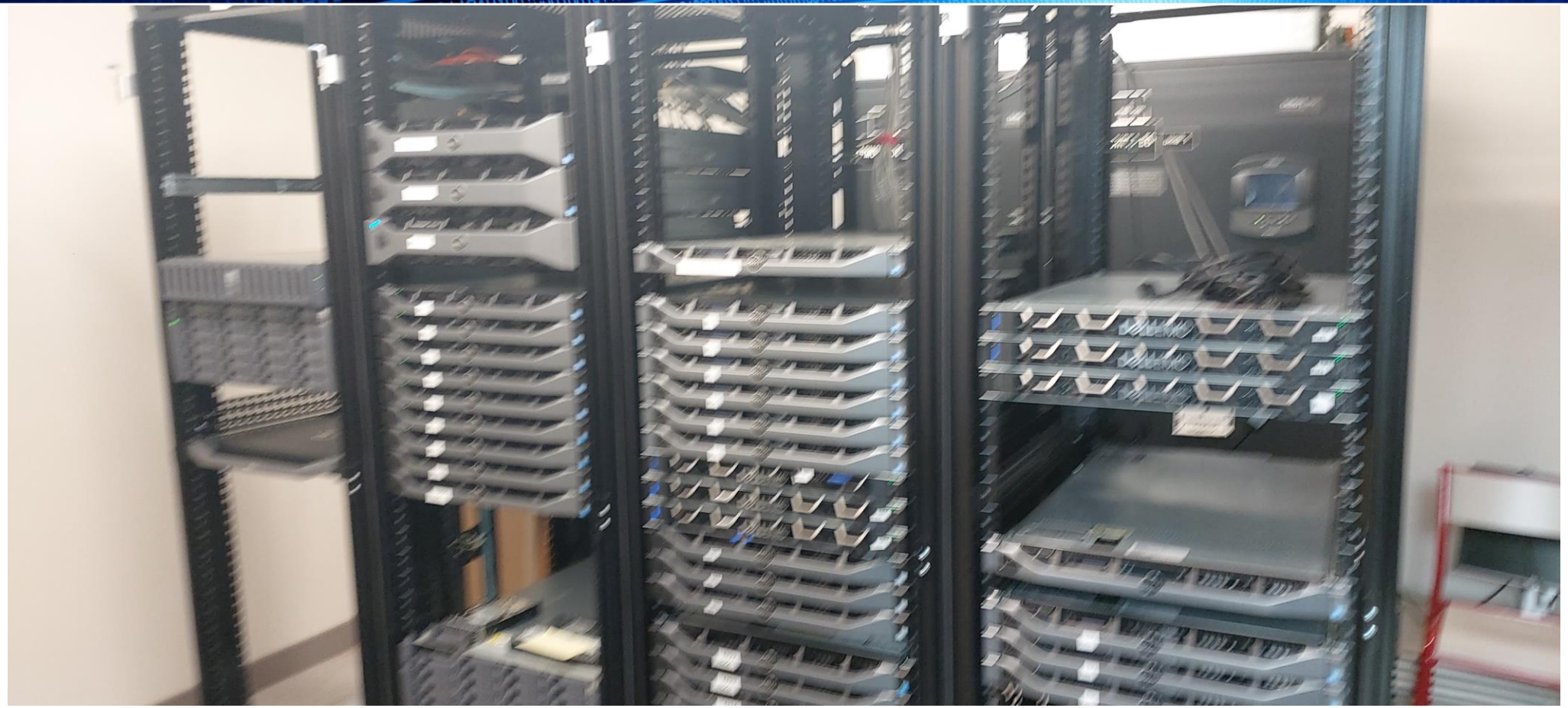
- Possibility of multiple domains -

Reviews

Small group photo

IRAP 2022 Barbecue

Reviews



TITAN

Reviews

- CNRS recruitment -
- CDD/CDI (external competitions) -
- Advice on preparing for competitive examinations-