NLP-chatbot R&D at CDS. VO standards, AI impact, etc.

André Schaaff, Thomas Boch, Sébastien Derriere

Centre de Données astronomiques de Strasbourg

IVOA Malta, Valetta Campus, 14-17 November 2024

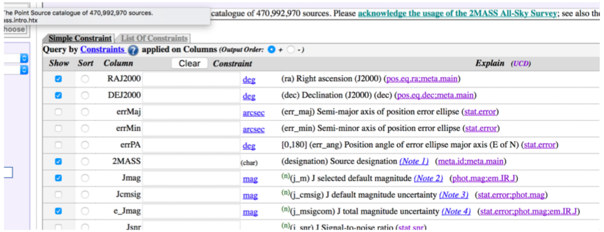
CENTRE DE DONNÉES ASTRONOMIQUES DE STRASBOURG

Purpose

- Feedback and status of a long term R&D action at CDS
- A study around Natural Language Processing applied to the astronomical data services querying
- It was presented step by step in various IVOA groups
- Taking now into account the AI vawe…

Why NLP ?

- Since Web advent we are used to deal with webpages containing forms (parameter fields, checkboxes, etc.)



he Point Source catalogue of 470,992,970 sources.
ss.intro.htx

hoose

catalogue of 470,992,970 sources. Please acknowledge the usage of the 2MASS All-Sky Survey; see also th

Simple Constraint | List Of Constraints

Query by Constraints ? applied on Columns *(Output Order:* ● + ○ - *)*

Show	*Sort*	*Column*	Clear	*Constraint*	*Explain (UCD)*
☑	○	RAJ2000		deg	(ra) Right ascension (J2000) (pos.eq.ra;meta.main)
☑	○	DEJ2000		deg	(dec) Declination (J2000) (dec) (pos.eq.dec;meta.main)
☐	○	errMaj		arcsec	(err_maj) Semi-major axis of position error ellipse (stat.error)
☐	○	errMin		arcsec	(err_min) Semi-minor axis of position error ellipse (stat.error)
☐	○	errPA		deg	[0,180] (err_ang) Position angle of error ellipse major axis (E of N) (stat.error)
☑	○	2MASS		(char)	(designation) Source designation *(Note 1)* (meta.id;meta.main)
☑	○	Jmag		mag	(n)(j_m) J selected default magnitude *(Note 2)* (phot.mag;em.IR.J)
☐	○	Jcmsig		mag	(n)(j_cmsig) J default magnitude uncertainty *(Note 3)* (stat.error;phot.mag)
☑	○	e_Jmag		mag	(n)(j_msigcom) J total magnitude uncertainty *(Note 4)* (stat.error;phot.mag;em.IR.J)
☐	○	Jsnr			(n)(j_snr) J Signal-to-noise ratio (stat.snr)

- Nowadays we use the voice (gesture should join for multimodal requests) to interact with several devices with voice answer or a displaying

- An interesting field of investigation with both technical and human concerns

Webpages / Forms pro cons

- Forms are not (yet) old fashion and have still advantages
  - Easy and quick for experimented users
  - easy to translate to database queries
  - Result rendering “easy” to design

- On the negative side :
  - not intuitive for new or occasional users
  - Thus we use often only a (small) part of the functionalities

- Understanding / Knowledge / Learning efforts on the user side

☐ NLP pro cons

- A friendly (Inclusive if disabilities) way of interaction
  - More digestible than forms
  - Just a prompt or a voice recognition
  - Result displaying can be similar than for forms or a speech synthesis

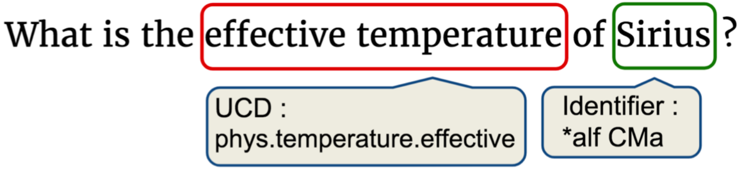
- On the negative side :
  - possible frustration for expert users if the result is too imprecise
  - too long chat if the question was vague or ambiguous

- User request Understanding and translation to query the services is on the service side

Pragmatic and best effort approach

- Learning about NLP (basis, tools, examples, …)
- Starting by building a set of typical queries

- We had already a lot of helpful data and tools
  - authors in Simbad, VizieR
  - missions and wavelengths in VizieR
  - DJIN to recognize identifiers in a text
  - UCDs
  - a name resolver

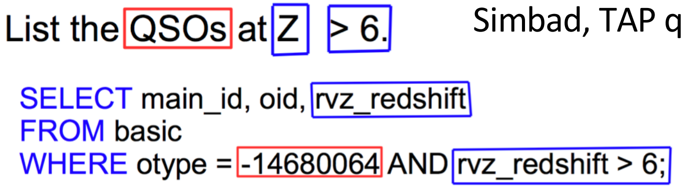

What is the effective temperature of Sirius ?

UCD : phys.temperature.effective

Identifier : *alf CMa

And a useful friend : the Virtual Observatory

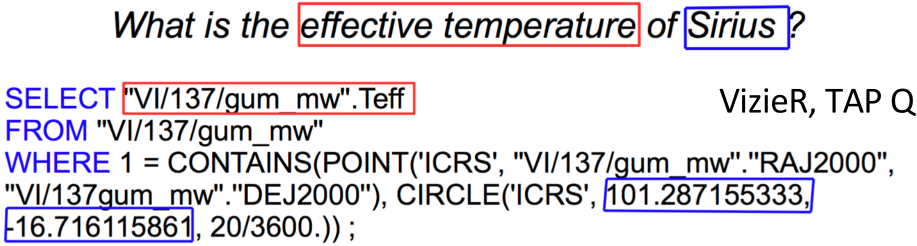
- We used VO standards like TAP/ADQL (DAL WG), UCDs (Semantics WG), … ,and tools based on ML (and now AI), a link with KDD IG.
- The VO was a master piece which helped us to translate NL into queries to our own services. By the way all services implementing VO standards could be included.



List the QSOs at Z > 6.

```
SELECT main_id, oid, rvz_redshift
FROM basic
WHERE otype = -14680064 AND rvz_redshift > 6;
```

Simbad, TAP query



What is the effective temperature of Sirius ?

```
SELECT "VI/137/gum_mw".Teff
FROM "VI/137/gum_mw"
WHERE 1 = CONTAINS(POINT('ICRS', "VI/137/gum_mw"."RAJ2000",
"VI/137gum_mw"."DEJ2000"), CIRCLE('ICRS', 101.287155333,
-16.716115861, 20/3600.)) ;
```

VizieR, TAP Query

☐ On the way to a chatbot

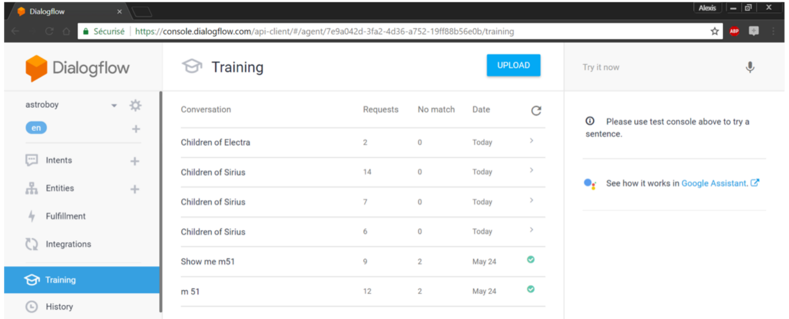
- A first work with the Stanford NLP library was far from a Chatbot as it must be able to chat with the user otherwise it is just like a command line

- Google Dialogflow brought us this chat capability
  - Based on Machine Learning
  - Training, tuning, improving features

Dialogflow

- Query understanding was then delegated to Dialogflow

Switching from coding to training to understand user intents

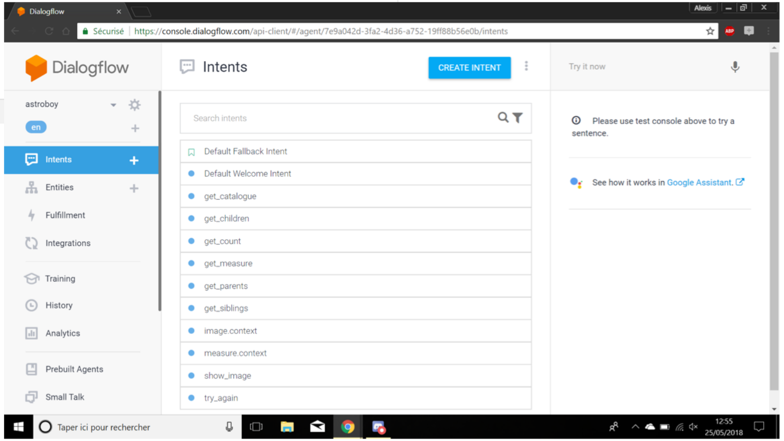


https://console.dialogflow.com/api-client/#/agent/7e9a042d-3fa2-4d36-a752-19ff88b56e0b/training

Conversation	Requests	No match	Date
Children of Electra	2	0	Today
Children of Sirius	14	0	Today
Children of Sirius	7	0	Today
Children of Sirius	6	0	Today
Show me m51	9	2	May 24
m 51	12	2	May 24

Please use test console above to try a sentence.

See how it works in Google Assistant.

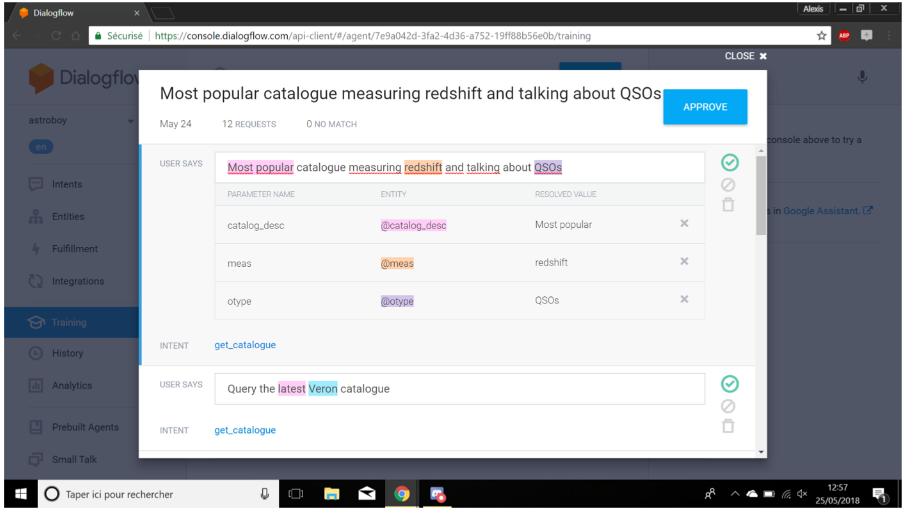


https://console.dialogflow.com/api-client/#/agent/7e9a042d-3fa2-4d36-a752-19ff88b56e0b/intents

- Default Fallback Intent
- Default Welcome Intent
- get_catalogue
- get_children
- get_count
- get_measure
- get_parents
- get_siblings
- image.context
- measure.context
- show_image
- try_again

Please use test console above to try a sentence.

See how it works in Google Assistant.



https://console.dialogflow.com/api-client/#/agent/7e9a042d-3fa2-4d36-a752-19ff88b56e0b/training

Most popular catalogue measuring redshift and talking about QSOs

May 24 12 REQUESTS 0 NO MATCH

USER SAYS: Most popular catalogue measuring redshift and talking about QSOs

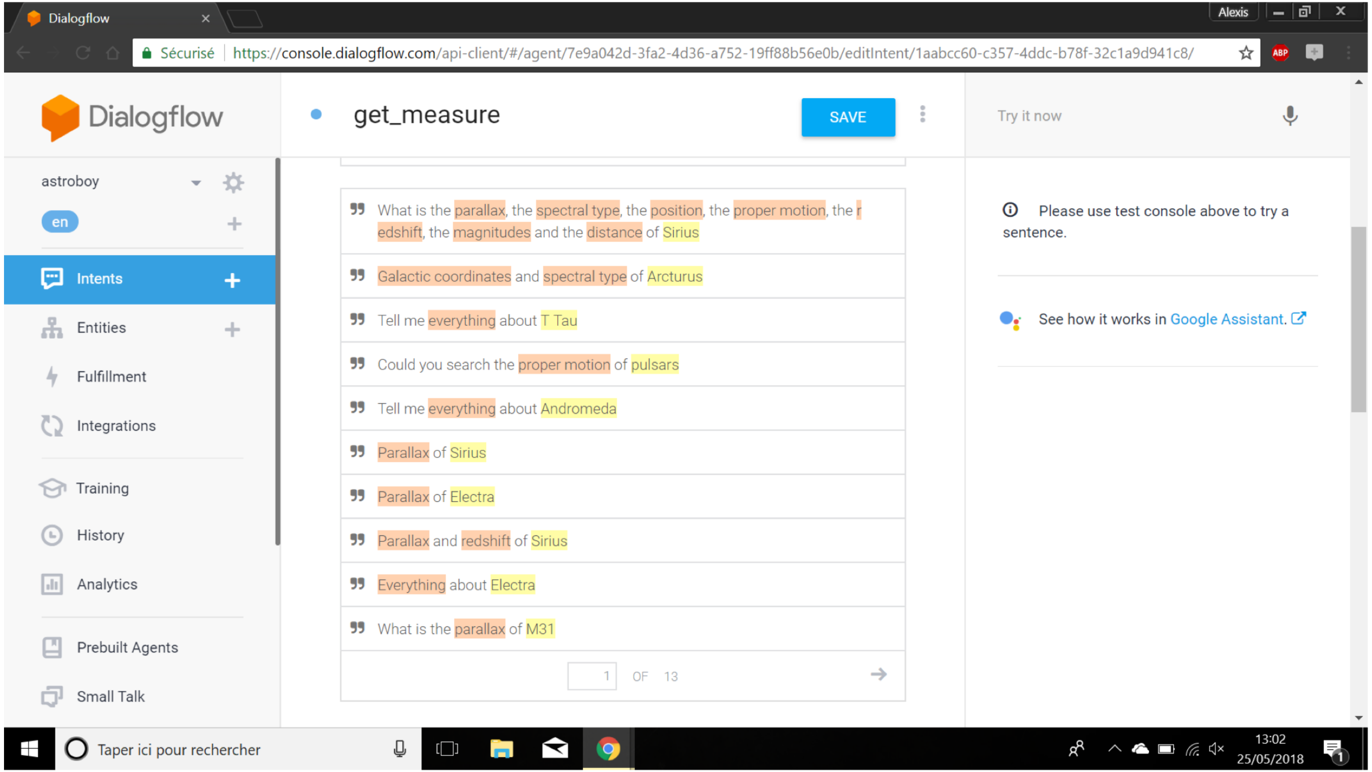
PARAMETER NAME	ENTITY	RESOLVED VALUE
catalog_desc	@catalog_desc	Most popular
meas	@meas	redshift
otype	@otype	QSOs

INTENT: get_catalogue

USER SAYS: Query the latest Veron catalogue

INTENT: get_catalogue

Examples of sentences linked to get_measure



- ❞ What is the parallax, the spectral type, the position, the proper motion, the redshift, the magnitudes and the distance of Sirius
- ❞ Galactic coordinates and spectral type of Arcturus
- ❞ Tell me everything about T Tau
- ❞ Could you search the proper motion of pulsars
- ❞ Tell me everything about Andromeda
- ❞ Parallax of Sirius
- ❞ Parallax of Electra
- ❞ Parallax and redshift of Sirius
- ❞ Everything about Electra
- ❞ What is the parallax of M31

ⓘ Please use test console above to try a sentence.

See how it works in Google Assistant.

Remark: External tools

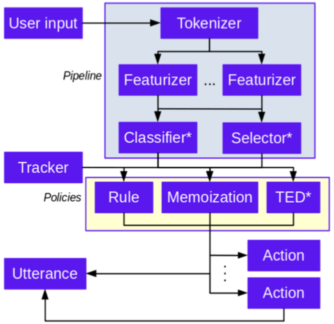
- With Dialogflow we delegated a larger part of the tool than with the first approach but not the link to the services.
- The link to the services and de facto to the data is the key part and our development was always designed to switch to other NLP tools.
- We switched to an Open Source framework, Rasa

☐ How does Rasa work ?

- … Rasa prediction model training is based on Tensorflow

- Compared to Dialogflow the configuration is not really user-friendly

  – Items of the architecture have to be written
  – For example a CDSTokeniser

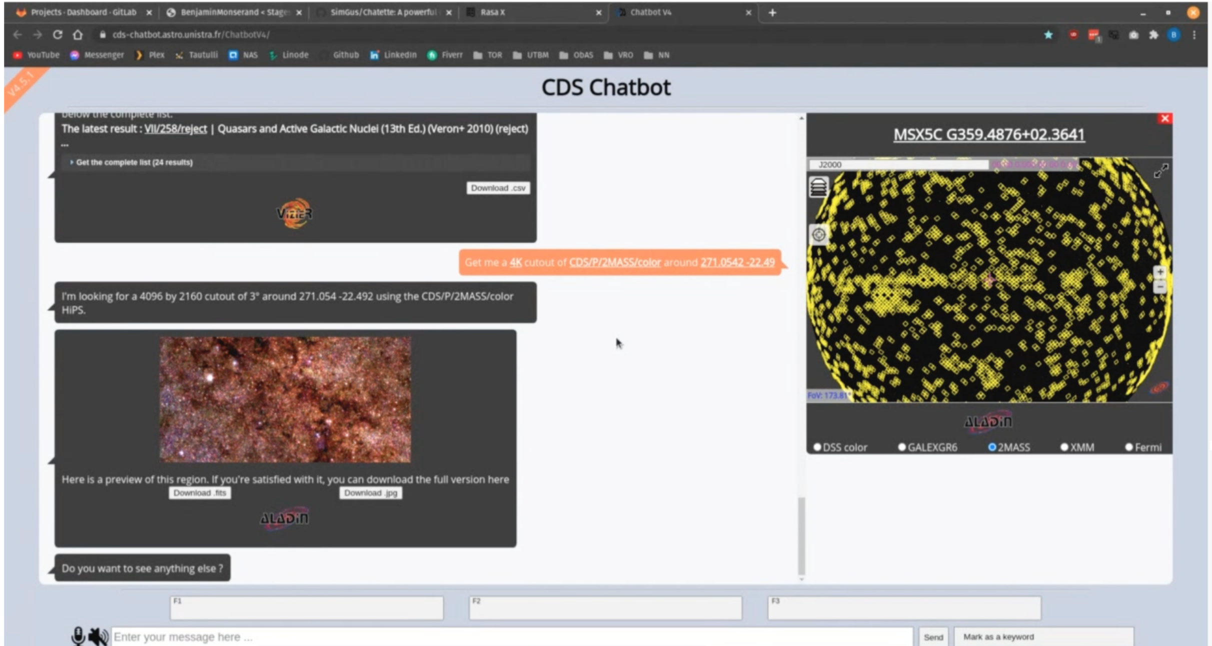


```
language: "en"

pipeline:
  - name: CDSTokenizer
  - name: RegexFeaturizer
  - name: LexicalSyntacticFeaturizer
  - name: CountVectorsFeaturizer
    analyzer: "word"
  - name: CountVectorsFeaturizer
    analyzer: "char_wb"
    min_ngram: 1
    max_ngram: 4
  - name: DIETClassifier
    epochs: 100
  - name: ResponseSelector
    epochs: 100

policies:
  - name: MemoizationPolicy
    max_history: 5
  - name: TEDPolicy
    max_history: 5
    epochs: 200
  - name: RulePolicy
```

Current version based on Rasa framework



CDS Chatbot

The latest result : VII/258/reject | Quasars and Active Galactic Nuclei (13th Ed.) (Veron+ 2010) (reject)

...

Get the complete list (24 results)

Download .csv

Get me a 4K cutout of CDS/P/2MASS/color around 271.0542 -22.49

I'm looking for a 4096 by 2160 cutout of 3° around 271.054 -22.492 using the CDS/P/2MASS/color HiPS.

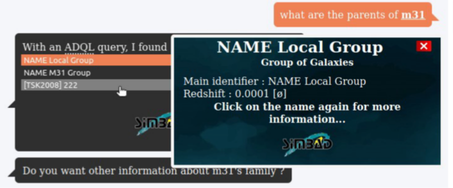
Here is a preview of this region. If you're satisfied with it, you can download the full version here

Download .fits Download .jpg

Do you want to see anything else ?

MSX5C G359.4876+02.3641

Enter your message here ...



what are the parents of m31

With an ADQL query, I found

NAME Local Group

NAME M31 Group

[TSK2008] 222

NAME Local Group

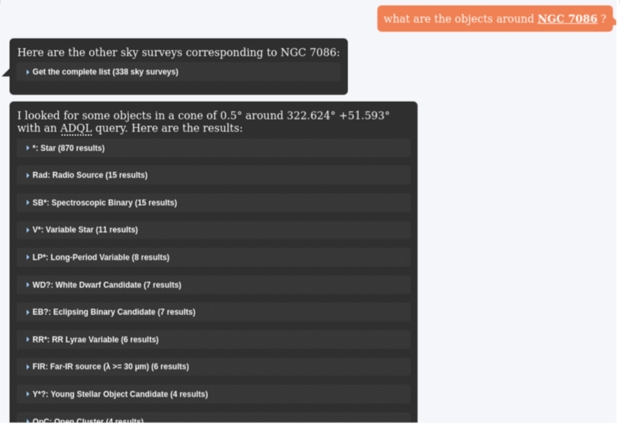
Group of Galaxies

Main identifier : NAME Local Group

Redshift : 0.0001 [ø]

Click on the name again for more information...

Do you want other information about m31's family ?



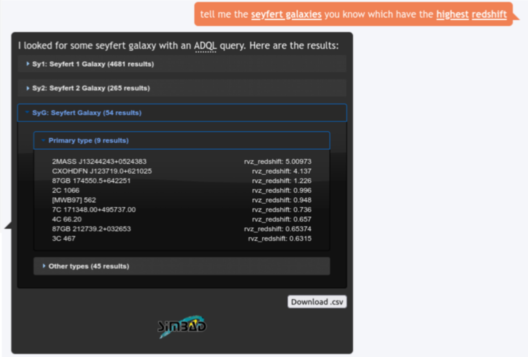
what are the objects around NGC 7086 ?

Here are the other sky surveys corresponding to NGC 7086:

- Get the complete list (338 sky surveys)

I looked for some objects in a cone of 0.5° around 322.624° +51.593° with an ADQL query. Here are the results:

- *: Star (870 results)
- Rad: Radio Source (15 results)
- SB*: Spectroscopic Binary (15 results)
- V*: Variable Star (11 results)
- LP*: Long-Period Variable (8 results)
- WD?: White Dwarf Candidate (7 results)
- EB?: Eclipsing Binary Candidate (7 results)
- RR*: RR Lyrae Variable (6 results)
- FIR: Far-IR source (λ >= 30 μm) (6 results)
- Y*?: Young Stellar Object Candidate (4 results)
- OpC: Open Cluster (4 results)



tell me the seyfert galaxies you know which have the highest redshift

I looked for some seyfert galaxy with an ADQL query. Here are the results:

- Sy1: Seyfert 1 Galaxy (4681 results)
- Sy2: Seyfert 2 Galaxy (265 results)
- SyG: Seyfert Galaxy (54 results)
  - Primary type (9 results)

2MASS J13244243+0524383	rvz_redshift: 5.00973
CXOHDFN J123719.0+621025	rvz_redshift: 4.137
87GB 174550.5+642251	rvz_redshift: 1.226
2C 1066	rvz_redshift: 0.996
[MWB97] 562	rvz_redshift: 0.948
7C 171348.00+495737.00	rvz_redshift: 0.736
4C 66.20	rvz_redshift: 0.657
87GB 212739.2+032653	rvz_redshift: 0.65374
3C 467	rvz_redshift: 0.6315

  - Other types (45 results)

Download .csv

The Day after

- Work with Rasa was continued (cleaning, refactoring, etc.) but with the advent of chatGPT we decided to make a break and to spend time on the study of its impact, a help for the existing chatbot ? or a way to reshape it completely ?

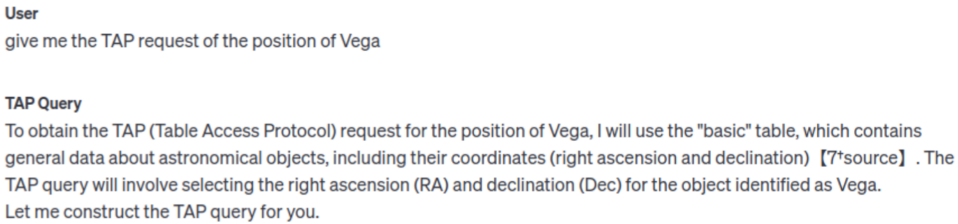
- We had a look at several AI tools and we decided to test with the OpenAI API which cost model (credits to buy and not a monthly subscription) is possible for us

First Experiments with OpenAI

- OpenAI has a functionnality called assistant which allows the definition of assistants inside an application, able to answser a request using models, tools ans data specified by the developper

- A first prototype was defined to learn Simbad table structure and ADQL to become able to propose or verify TAP queries to Simbad.

Examples

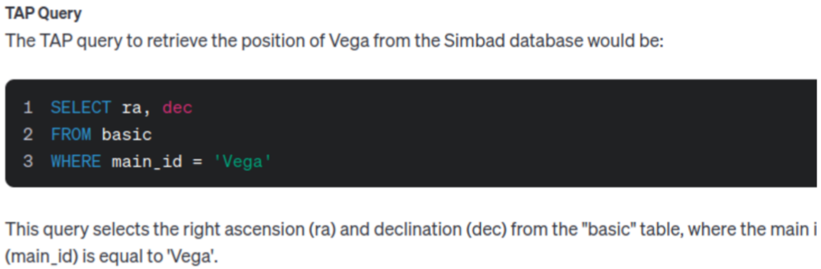


User

give me the TAP request of the position of Vega

TAP Query

To obtain the TAP (Table Access Protocol) request for the position of Vega, I will use the "basic" table, which contains general data about astronomical objects, including their coordinates (right ascension and declination) 【7†source】. The TAP query will involve selecting the right ascension (RA) and declination (Dec) for the object identified as Vega.
Let me construct the TAP query for you.

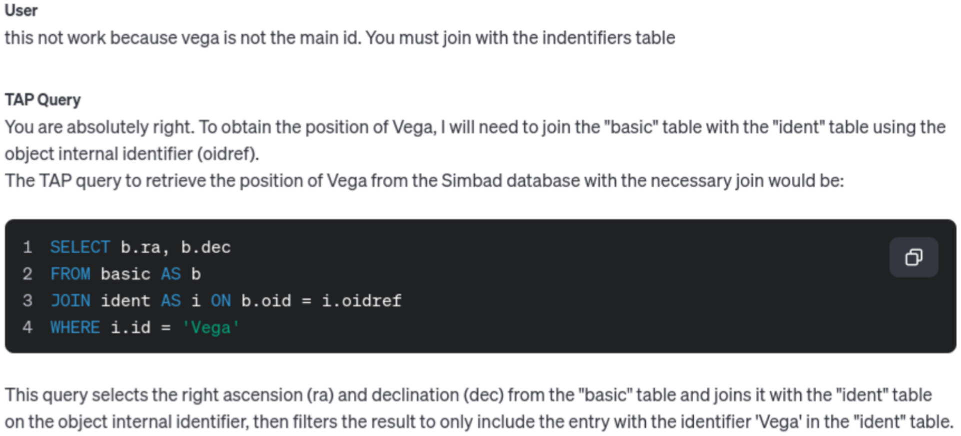


TAP Query

The TAP query to retrieve the position of Vega from the Simbad database would be:

```sql
SELECT ra, dec
FROM basic
WHERE main_id = 'Vega'
```

This query selects the right ascension (ra) and declination (dec) from the "basic" table, where the main i
(main_id) is equal to 'Vega'.



User

this not work because vega is not the main id. You must join with the indentifiers table

TAP Query

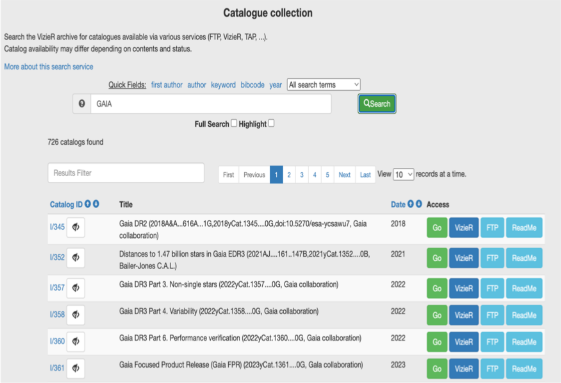
You are absolutely right. To obtain the position of Vega, I will need to join the "basic" table with the "ident" table using the object internal identifier (oidref).
The TAP query to retrieve the position of Vega from the Simbad database with the necessary join would be:

```sql
SELECT b.ra, b.dec
FROM basic AS b
JOIN ident AS i ON b.oid = i.oidref
WHERE i.id = 'Vega'
```

This query selects the right ascension (ra) and declination (dec) from the "basic" table and joins it with the "ident" table on the object internal identifier, then filters the result to only include the entry with the identifier 'Vega' in the "ident" table.

Ongoing experiments

- We have a mine of information not directly used in the queries to the services like the VizieR Readme files



Catalogue collection

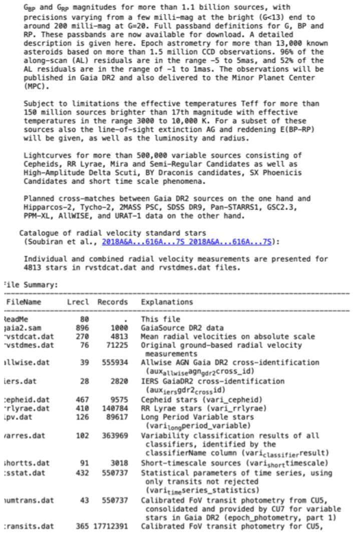
Search the VizieR archive for catalogues available via various services (FTP, VizieR, TAP, ...).
Catalog availability may differ depending on contents and status.

More about this search service
Quick Fields: first author author keyword bibcode year
All search terms
GAIA
Search
Full Search ☐ Highlight ☐

726 catalogs found

Results Filter
First Previous 1 2 3 4 5 Next Last
View 10 records at a time.

Catalog ID	Title	Date	Access
I/345	Gaia DR2 (2018A&A...616A...1G,2018yCat.1345....0G,doi:10.5270/esa-ycsawu7, Gaia collaboration)	2018	Go VizieR FTP ReadMe
I/352	Distances to 1.47 billion stars in Gaia EDR3 (2021AJ....161..147B,2021yCat.1352....0B, Bailer-Jones C.A.L.)	2021	Go VizieR FTP ReadMe
I/357	Gaia DR3 Part 3. Non-single stars (2022yCat.1357....0G, Gaia collaboration)	2022	Go VizieR FTP ReadMe
I/358	Gaia DR3 Part 4. Variability (2022yCat.1358....0G, Gaia collaboration)	2022	Go VizieR FTP ReadMe
I/360	Gaia DR3 Part 6. Performance verification (2022yCat.1360....0G, Gaia collaboration)	2022	Go VizieR FTP ReadMe
I/361	Gaia Focused Product Release (Gaia FPR) (2023yCat.1361....0G, Gala collaboration)	2023	Go VizieR FTP ReadMe



```
     GBP and GRP magnitudes for more than 1.1 billion sources, with
     precisions varying from a few milli-mag at the bright (G<13) end to
     around 200 milli-mag at G=20. Full passband definitions for G, BP and
     RP. These passbands are now available for download. A detailed
     description is given here. Epoch astrometry for more than 13,000 known
     asteroids based on more than 1.5 million CCD observations. 96% of the
     along-scan (AL) residuals are in the range -5 to 5mas, and 52% of the
     AL residuals are in the range of -1 to 1mas. The observations will be
     published in Gaia DR2 and also delivered to the Minor Planet Center
     (MPC).

     Subject to limitations the effective temperatures Teff for more than
     150 million sources brighter than 17th magnitude with effective
     temperatures in the range 3000 to 10,000 K. For a subset of these
     sources also the line-of-sight extinction AG and reddening E(BP-RP)
     will be given, as well as the luminosity and radius.

     Lightcurves for more than 500,000 variable sources consisting of
     Cepheids, RR Lyrae, Mira and Semi-Regular Candidates as well as
     High-Amplitude Delta Scuti, BY Draconis candidates, SX Phoenicis
     Candidates and short time scale phenomena.

     Planned cross-matches between Gaia DR2 sources on the one hand and
     Hipparcos-2, Tycho-2, 2MASS PSC, SDSS DR9, Pan-STARRS1, GSC2.3,
     PPM-XL, AllWISE, and URAT-1 data on the other hand.

    Catalogue of radial velocity standard stars
     (Soubiran et al., 2018A&A...616A...7S 2018A&A...616A...7S):

     Individual and combined radial velocity measurements are presented for
     4813 stars in rvstdcat.dat and rvstdmes.dat files.

ile Summary:
--------------------------------------------------------------------------------
 FileName      Lrecl  Records   Explanations
--------------------------------------------------------------------------------
eadMe            80        .   This file
aia2.sam        896     1000   GaiaSource DR2 data
vstdcat.dat     270     4813   Mean radial velocities on absolute scale
vstdmes.dat      76    71225   Original ground-based radial velocity
                                measurements
llwise.dat       39   555934   Allwise AGN Gaia DR2 cross-identification
                                (auxallwiseagngdr2cross_id)
ers.dat          28     2820   IERS GaiaDR2 cross-identification
                                (auxiersgdr2crossid)
epheid.dat      467     9575   Cepheid stars (vari_cepheid)
rlyrae.dat      410   140784   RR Lyrae stars (vari_rrlyrae)
pv.dat          126    89617   Long Period Variable stars
                                (varilongperiod_variable)
arres.dat       102   363969   Variability classification results of all
                                classifiers, identified by the
                                classifierName column (variclassifierresult)
hortts.dat       91     3018   Short-timescale sources (varishorttimescale)
sstat.dat       432   550737   Statistical parameters of time series, using
                                only transits not rejected
                                (varitimeseries_statistics)
umtrans.dat      43   550737   Calibrated FoV transit photometry from CU5,
                                consolidated and provided by CU7 for variable
                                stars in Gaia DR2 (epoch_photometry, part 1)
ransits.dat     365 17712391   Calibrated FoV transit photometry for CU5,
```

LLMs are efficient for textual content
The aim is to train with this additionnal data to detect the astronomical object names in the Readme files

Files of valid, invalid names to train, etc.

remark

- Testing / working with AI frameworks is often nearer to cooking than to programming
- best recipe to find
- Trying with different models
- Training with a good set of data (also a set with errors)
- Improving with other sources or various data formats for a better understanding by the models
- Etc.

Conclusion

- We are testing and improving step by step to interact with data services differently, keeping in mind the gap between a prototype and a production service

- .. And adapting us to new tools/frameworks/paradigms; so as usual in IT

Providing stable and sustainable new tools/services benefiting from AI is an interesting challenge

Questions ?