R&D @ CDS

and other developments

André Schaaff on behalf of the CDS Team

CDS Scientific Council 2017

Why R&D ?

- Technological evolutions are fast and in various IT fields (interactions, visualization, mobility, components, Big Data & Open data, Clouds, etc.) with many actors in both the commercial and the Open Source domains

- The CDS team has always spent time in technology watch to follow the evolutions but it is becoming hard to test and evaluate everything in addition to the everyday work

- The R&D activity is now well identifed, structured and involves several persons of the staff with the help of interns and short contracts

- It is a continous training of the IT team and it provides also inputs to present and discuss during the Infusion meetings

□ Internship programme

- 11 interns hired in 2017 to work with us on several topics, R&D and other developments
- A total of 2,5 years of internships per year…
- + short contracts
  - to push the work to the production side
  - to work on short developments during the Summer
- Tight IT Job Market => possible future hiring on projects

Internships & short contracts in 2017

- Evaluation of NoSQL technologies for Simbad criteria query
- PostgreSQL investigation for massive data in astronomy
- Evolution of Xfits, a tool dedicated to images and spectrum
- Natural Language Processing to request astronomical services
- 3D visualization in a Web browser (large datasets, interpretation and immersion)
- IVOA Provenance Model implementation for distributed databases

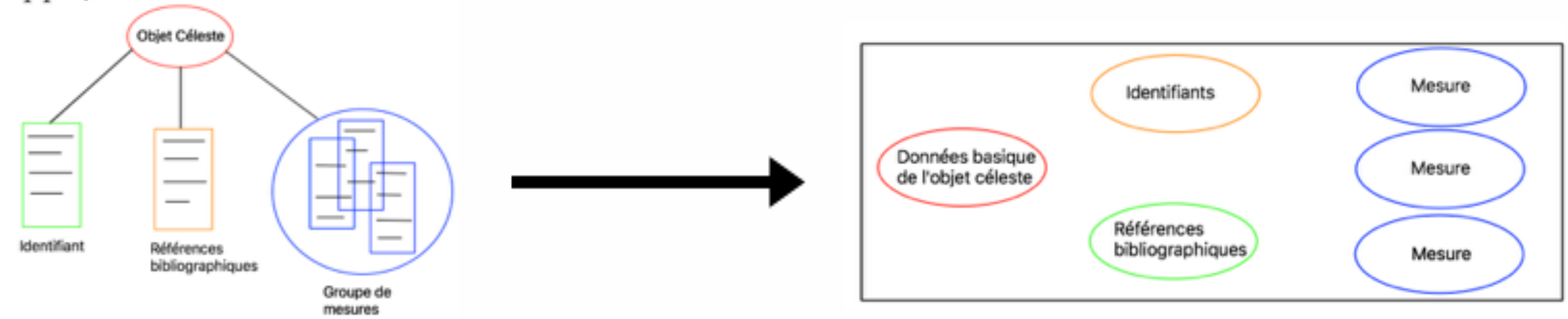
Internships & short contracts in 2017 (2)

- IVOA VOSpace API in Python Code near the data, Apache Spark / X-Match → on going
- Code near the data - Apache Spark and X-Match
- Prototype of a Jupyter notebooks server attached to CDS accounts
- Aladin Lite interface extension L. Michel (SSC XMM-Newton) → on going
- Video tutorials

Evaluation of NoSQL technologies for Simbad criteria query

- 3 technologies tested: Cassandra, cstore_fdw (PostgreSQL columnar store extension), ElasticSearch
  - Installation
  - Ingesting Simbad data
  - Benchmarks on a set of typical queries

- Requires a denormalization of Simbad DB schema



Objet Céleste

Identifiant

Références bibliographiques

Groupe de mesures

Données basique de l'objet céleste

Identifiants

Références bibliographiques

Mesure

Mesure

Mesure

- Conclusion: ElasticSearch shows promising results for queries on predefined fields

T. Boch, A. Oberto
Intern: Alexandre Sevin (IUT Dijon)

PostgreSQL investigation for massive data in astronomy

- For TAPVizieR and Simbad

- Exploring database technology for replication:
  - Buccardo (master-master replication)
  - Pgpool (replication , pool & load balancing)
  - Greenplum (parallel data arcitecture)

- Interesting technologies in the Big Data context
  - not used today because it increases maintenance and requires more in-depth knowledge

G. Landais, A. Oberto
Intern: Alexandre Vaquembergue (IUT Charlemagne, Nancy)

Complement in the frame of a short contract

- VizieR database into a Docker container
- Prototype of the new TAPVizieR architecture with HAProxy
- Exploration of the new PostgreSQL replication technology: Postgres-BDR (2ndQuadrant) based on pglogical replication

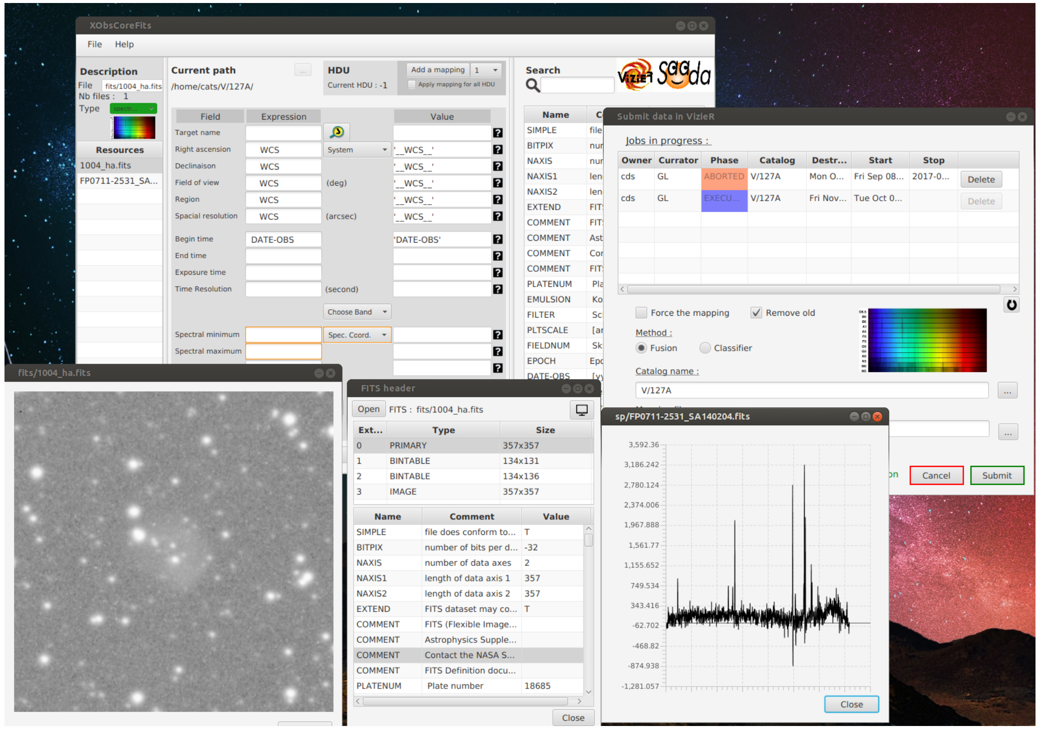
2ndQuadrant®
PostgreSQL

- => promising technology but not yet used in TAPVizieR as it is not free for PostgreSQL 9.6

G. Landais
Contractor: Paul Tréhiou

Evolution of XobsCoreFits, a tool dedicated to images and spectrum

- Dedicated functionality for the documentalists to make the mapping between FITS resources and the IVOA ObsCore Data Model
  - Spectra and images preview
  - Pilot the ingestion into the VizieR associated database



Description
File fits/1004_ha.fits
Nb files : 1
Type spectr...

Resources
1004_ha.fits
FP0711-2531_SA...

Current path
/home/cats/V/127A/

HDU
Current HDU : -1
☐ Apply mapping for all HDU

Field	Expression		Value
Target name			
Right ascension	WCS	System	'__WCS__'
Declinaison	WCS		'__WCS__'
Field of view	WCS	(deg)	'__WCS__'
Region	WCS		'__WCS__'
Spacial resolution	WCS	(arcsec)	'__WCS__'
Begin time	DATE-OBS		'DATE-OBS'
End time			
Exposure time			
Time Resolution		(second)	
		Choose Band	
Spectral minimum		Spec. Coord.	
Spectral maximum			

Name	C
SIMPLE	file
BITPIX	nur
NAXIS	nur
NAXIS1	len
NAXIS2	len
EXTEND	FIT
COMMENT	FIT
COMMENT	Ast
COMMENT	Cor
COMMENT	FIT
PLATENUM	Pla
EMULSION	Ko
FILTER	Sc
PLTSCALE	[ar
FIELDNUM	Sk
EPOCH	Ep
DATE-OBS	[yy

fits/1004_ha.fits

FITS header

Open FITS : fits/1004_ha.fits

Ext...	Type	Size
0	PRIMARY	357x357
1	BINTABLE	134x131
2	BINTABLE	134x136
3	IMAGE	357x357

Name	Comment	Value
SIMPLE	file does conform to...	T
BITPIX	number of bits per d...	-32
NAXIS	number of data axes	2
NAXIS1	length of data axis 1	357
NAXIS2	length of data axis 2	357
EXTEND	FITS dataset may co...	T
COMMENT	FITS (Flexible Image...	
COMMENT	Astrophysics Supple...	
COMMENT	Contact the NASA S...	
COMMENT	FITS Definition docu...	
PLATENUM	Plate number	18685

Submit data in VizieR

Jobs in progress :

Owner	Currator	Phase	Catalog	Destr...	Start	Stop	
cds	GL	ABORTED	V/127A	Mon O...	Fri Sep 08...	2017-0...	Delete
cds	GL	EXECU...	V/127A	Fri Nov...	Tue Oct 0...		Delete

☐ Force the mapping ☑ Remove old

Method :
● Fusion ○ Classifier

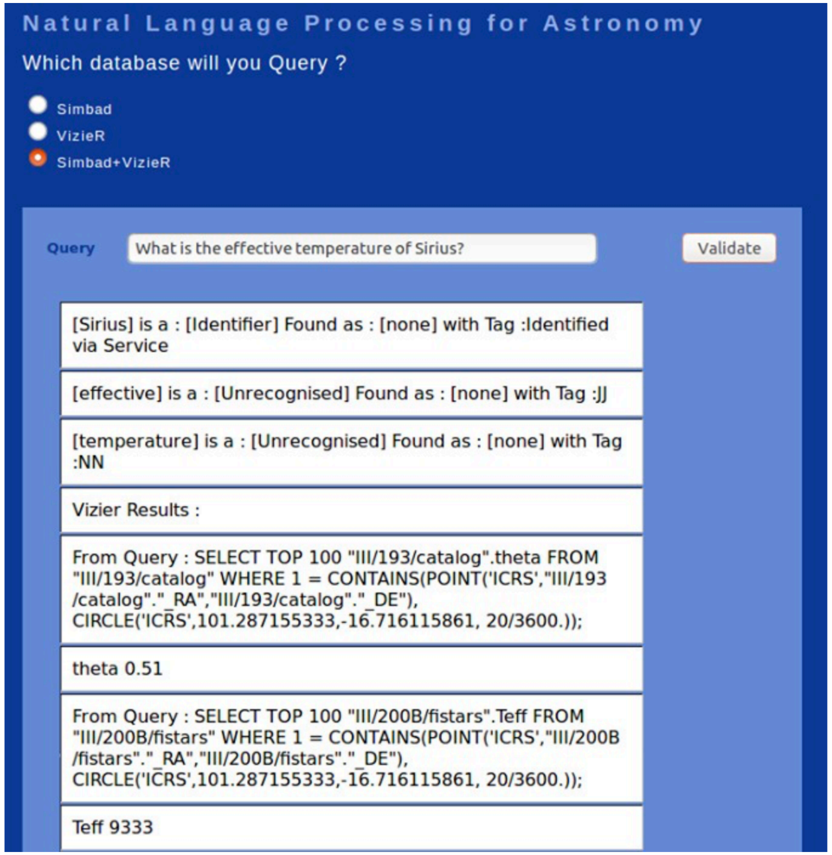
Catalog name :
V/127A

G. Landais
Intern: Félix Royer (L1, Université de Franche-Comté)

Natural Language Processing to request astronomical services

What is the effective temperature of Sirius ?

- Voice usage is becoming natural (Siri, Ok Google)
- An alternative to the current way based on forms (parameter fields, checkboxes, etc.) through a unique text field or a voice recognition of its content
- Based on Stanford NLP, DJIN, IVOA UCD and ADQL/TAP, …



Natural Language Processing for Astronomy

Which database will you Query ?

○ Simbad
○ VizieR
● Simbad+VizieR

Query: What is the effective temperature of Sirius?

```
[Sirius] is a : [Identifier] Found as : [none] with Tag :Identified
via Service

[effective] is a : [Unrecognised] Found as : [none] with Tag :JJ

[temperature] is a : [Unrecognised] Found as : [none] with Tag
:NN

Vizier Results :

From Query : SELECT TOP 100 "III/193/catalog".theta FROM
"III/193/catalog" WHERE 1 = CONTAINS(POINT('ICRS',"III/193
/catalog"."_RA","III/193/catalog"."_DE"),
CIRCLE('ICRS',101.287155333,-16.716115861, 20/3600.));

theta 0.51

From Query : SELECT TOP 100 "III/200B/fistars".Teff FROM
"III/200B/fistars" WHERE 1 = CONTAINS(POINT('ICRS',"III/200B
/fistars"."_RA","III/200B/fistars"."_DE"),
CIRCLE('ICRS',101.287155333,-16.716115861, 20/3600.));

Teff 9333
```

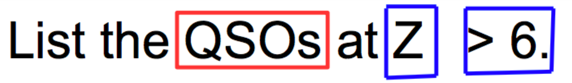
What is the redshift of galaxies members of the Virgo cluster ?

A. Schaaff, T. Boch, S. Derriere
Interns: Aymon De Saint-Acheul (IUT Charlemagne, Nancy), Pierre Barjon (ENSIIE Strasbourg)

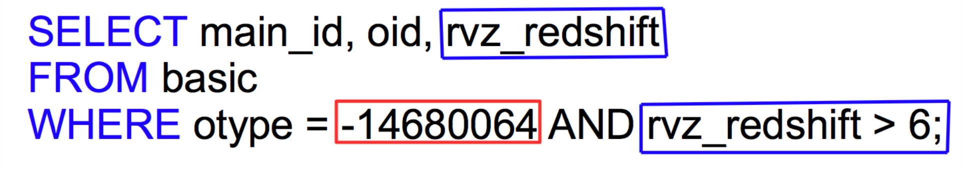
How many planets orbit Kepler 20 ?

Natural Language Processing to request astronomical services (2)

- Examples

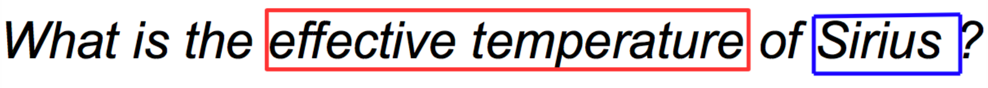


List the QSOs at Z > 6.

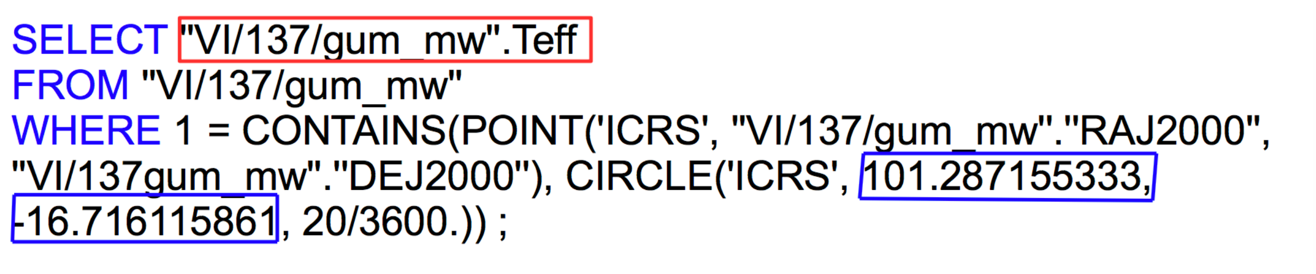


```
SELECT main_id, oid, rvz_redshift
FROM basic
WHERE otype = -14680064 AND rvz_redshift > 6;
```

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

A. Schaaff, T. Boch, S. Derriere
Interns: Aymon De Saint-Acheul (IUT Charlemagne, Nancy), Pierre Barjon (ENSIIE Strasbourg)

□ 3D Visualization in a Web browser

- Light tool to visualize several kinds of 3D data in a Web browser (based on WebGL)

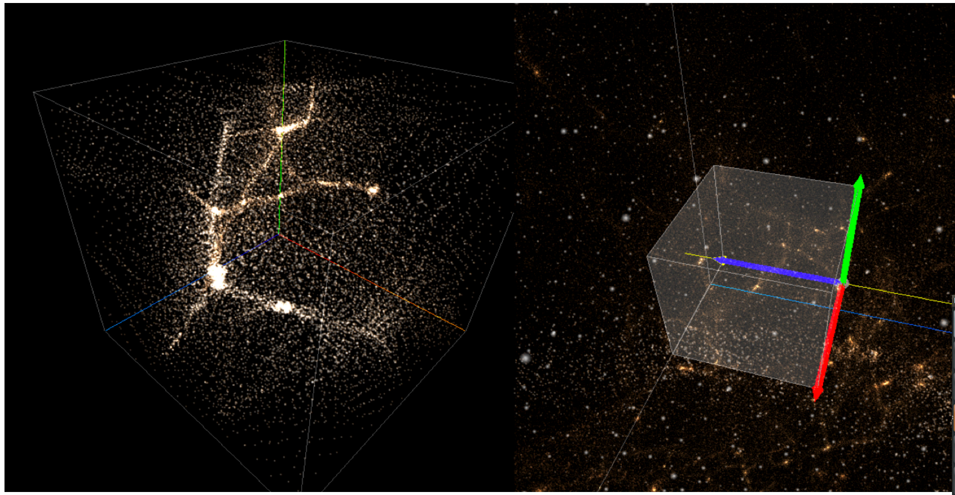
- Since R&D 2016 a work was on the server side to enable the visualization of large datasets:
  - 4096^3 simulation data cube (a few TBs)
  - data on a server + progressive visualization on the client side ("à la HiPS" but for cubes with all-directions visualization)

- Paper for A&C in preparation for R&D 2015-2017

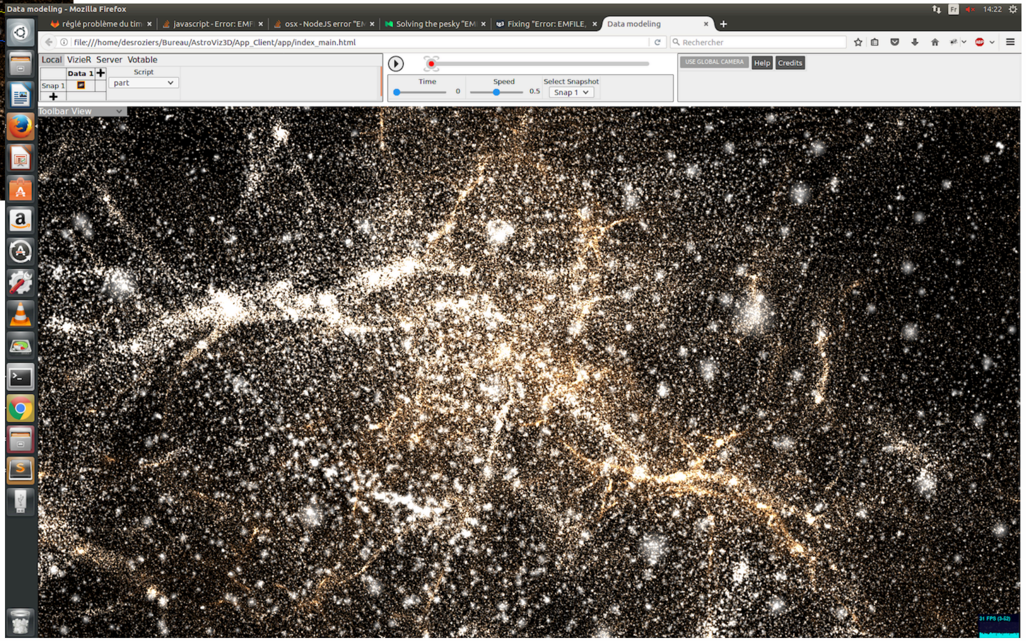
A. Schaaff, D. Aubert, N. Deparis, N. Gillet, P. Ocvirk , F.-X. Pineau
Interns: Malek El Ouerghi (ENSIIE Strasbourg), Jérôme Desroziers (Telecom Nancy)

3D Visualization in a Web browser (illustration 1)

Navigation in the data



Emma Simulation Data


Data modeling - Mozilla Firefox

file:///home/desroziers/Bureau/AstroViz3D/App_Client/app/index_main.html

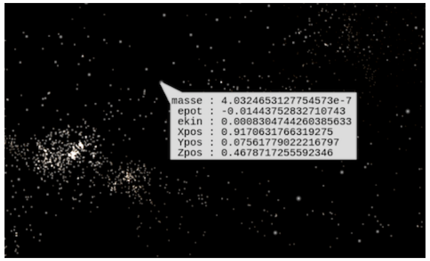
Script: part

Time 0

Speed 0.5

Select Snapshot: Snap 1

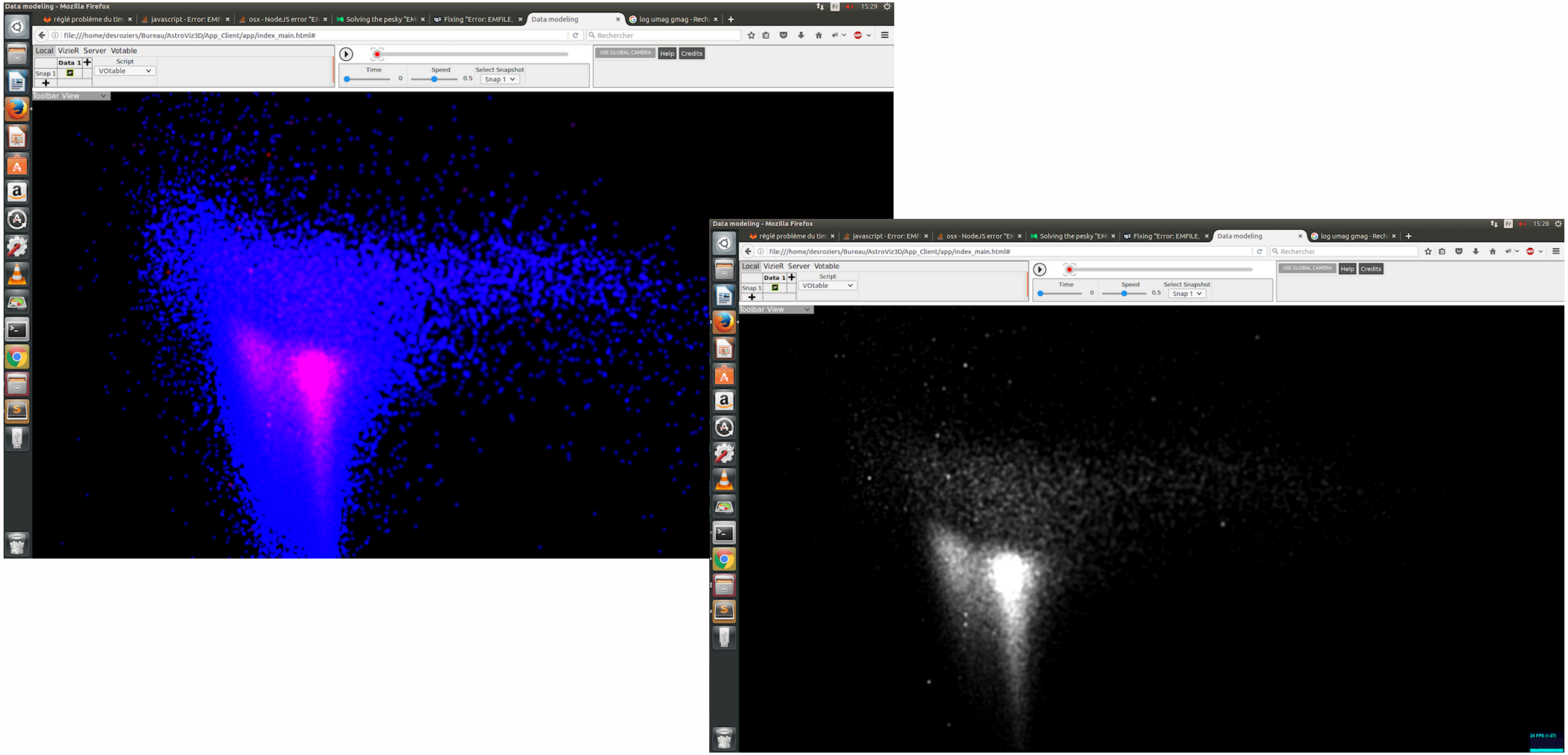
Source file information



```
masse : 4.0324653127754573e-7
 epot : -0.01443752832710743
 ekin : 0.0008304744260385633
 Xpos : 0.9170631766319275
 Ypos : 0.07561779022216797
 Zpos : 0.4678717255592346
```

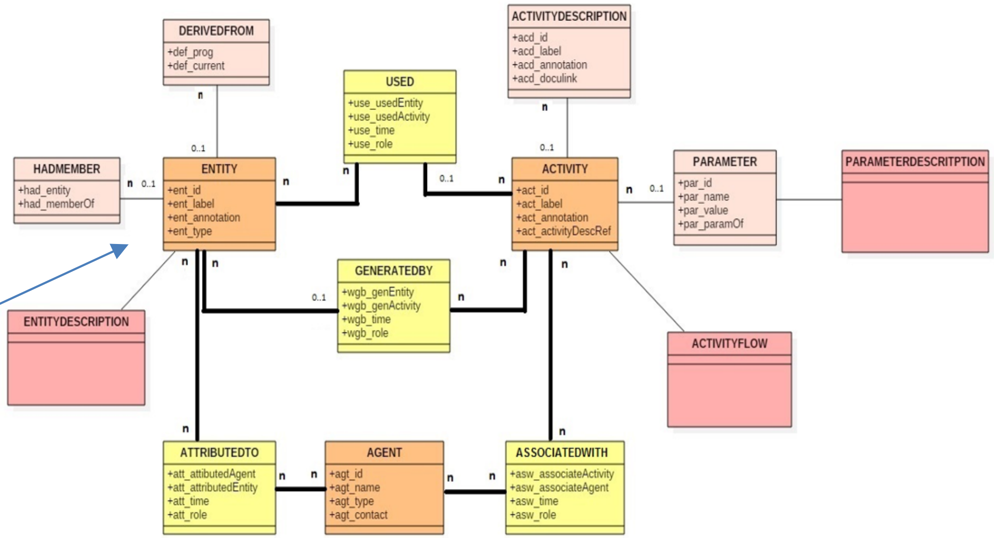
3D Visualization in a Web browser (illustration 2)

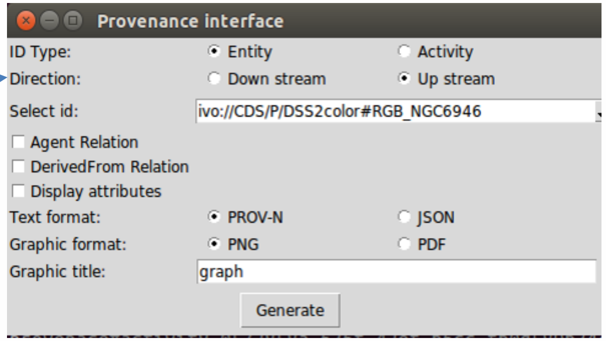
Not dedicated to simulation data, an example with the VizieR SDSS catalogue



IVOA Provenance Model implementation for distributed databases

- The goal is to track Provenance metadata of image datasets for colour composition, cut-outs, plate digitization and Hips generation
  - Mapping of the Provenance UML model into a relational database
  - Importing / exporting in PROV-N(W3C), JSON and VOTable in a specific PROV-VOTable document template
  - Compatibility of responses with VO tools (TopCat, …)
  - User interface and W3C graphs outputs





Provenance interface

ID Type: ● Entity ○ Activity

Direction: ○ Down stream ● Up stream

Select id: ivo://CDS/P/DSS2color#RGB_NGC6946

□ Agent Relation

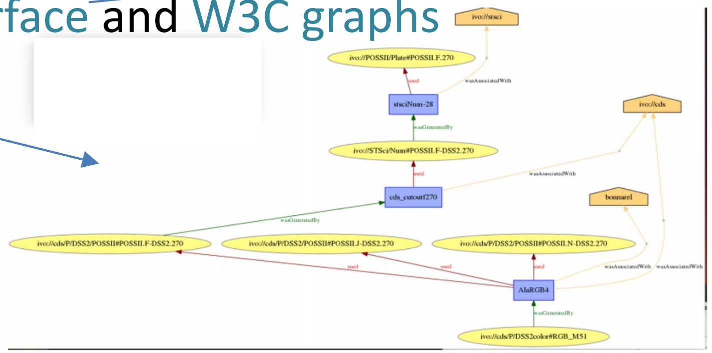
□ DerivedFrom Relation

□ Display attributes

Text format: ● PROV-N ○ JSON

Graphic format: ● PNG ○ PDF

Graphic title: graph


ivo://stsci
ivo://POSSII/Plate#POSSII.F.270
used
wasAssociatedWith
stsciNum-28
wasGeneratedBy
ivo://cds
ivo://STSci/Num#POSSII.F-DSS2.270
used
wasAssociatedWith
cds_cutoutf270
bonnarel
wasGeneratedBy
ivo://cds/P/DSS2/POSSII#POSSII.F-DSS2.270
ivo://cds/P/DSS2/POSSII#POSSII.J-DSS2.270
ivo://cds/P/DSS2/POSSII#POSSII.N-DSS2.270
used
used
used
wasAssociatedWith
wasAssociatedWith
AlaRGB4
wasGeneratedBy
ivo://cds/P/DSS2color#RGB_M51

M. Louys, F. Bonnarel
Intern: François Bock (IUT Schuman, Strasbourg)

□ IVOA VOSpace API in Python

- VOSpace is the IVOA protocol to access storage systems, as an overlay
- Implementation is not easy
- Existing implementation are not often fully compliant
- The aim is to have our own implementation to test our tools and to provide light VOSpace overlays

A. Schaaff, I. Yapici, T. Boch, P. Fernique
Intern: Madjid Bouchair (LP, Université de Haute Alsace)

Code near the data - Apache Spark and X-Match

- Apache Spark evaluated in the frame of a use case, the "cross-match" of source catalogues use case (presented at several occasions, collaborations, etc.)
- On going work with a focus on how (framework, security, hardware & software needs/costs) to bring the code to the data -> needs of large projects
- Maybe a paper in A&C

A. Schaaff, F.-X. Pineau, O. Aidel (IN2P3), J. Nauroy (Paris-Sud), T. Boch, G. Landais, L. Michel
Interns: Corentin Sanchez (UTBM) (Paul Trehiou (UTBM) in 2016/17, Noémie Wali (UTBM) in 2015/16)

Prototype of a Jupyter notebooks server attached to CDS accounts

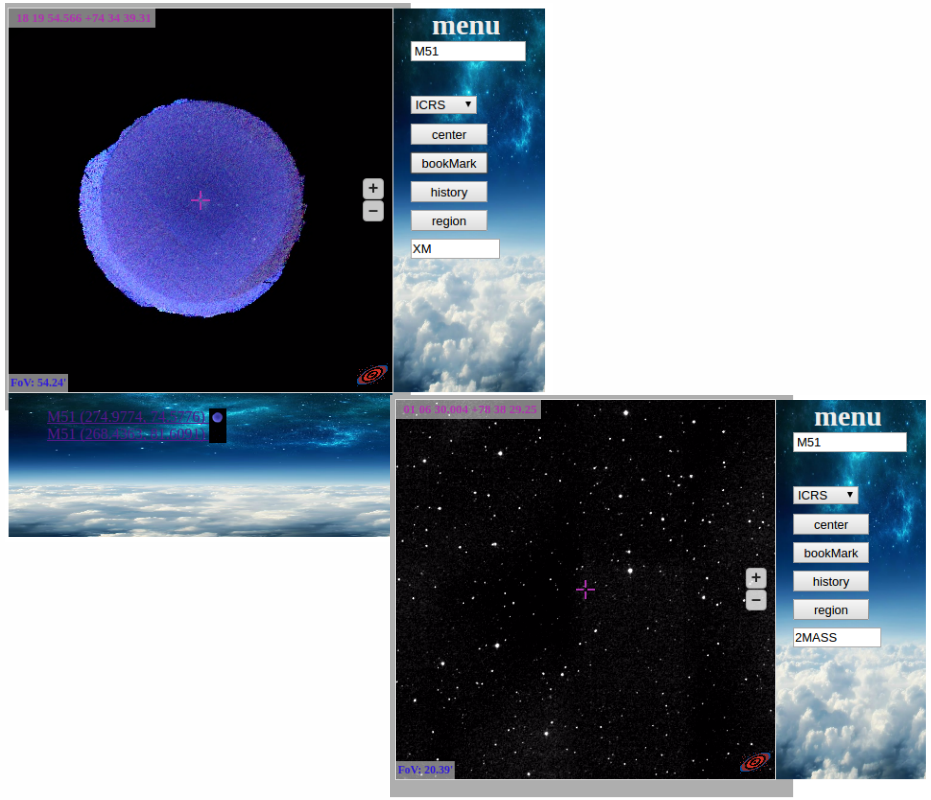
- At login a Docker container is started, containing:
  - CDS tuned Jupyter notebook (astropy, Aladin Lite plugin, ... are preloaded)
  - Volume with a limited space mounted for user usage with previously saved data / scripts
  - Access to data stored in MyCDS from Jupyter

- Managing of the security aspects (limited CPU resources and rights (not root) to limit the impact in case of hacking

F.-X. Pineau, T. Boch for the AladinLite plugin
Contractors: Paul Tréhiou, Jérôme Desroziers for the AladinLite plugin

□ Aladin Lite interface extension

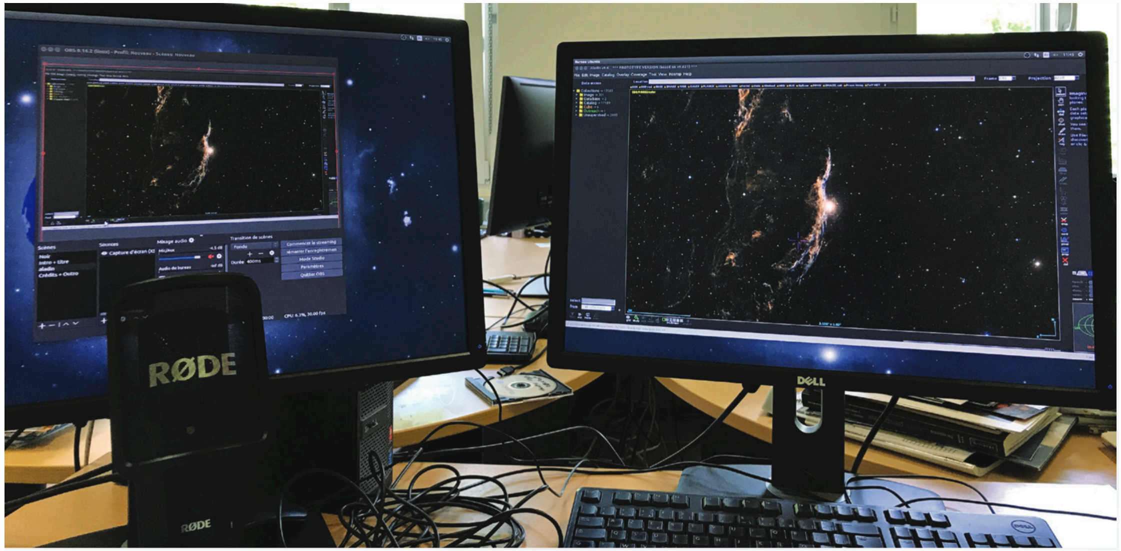
- External retractable panels
- Text search on the MOC server (HiPS and VizieR tables, resources selection constrained on the FoV)
- History (view storage / annotation)
- Polygonal regions tracing (catalogue data selection, interface to query by regions)
- Connexion to TAP



L. Michel (SSC XMM-Newton), T. Boch, A. Schaaff
Intern: Jie Wang (UTBM Belfort-Montbéliard)

□ Video tutorials

- The aim was to define how to produce as easily as possible video tutorials for the CDS services and tools
- Example



S. Derriere, A. Schaaff
Intern: Cédric Vogel (IUT Saint-Dié-des-Vosges)

The studio…

A selection of News since R&D 2016

- Google Summer of Code (Thomas): a HiPS Python library development

- Amazon AWS Research Credits rewards (André, Thomas, François-Xavier, Anais, Pierre): in the clouds tests for HiPS generation / distribution, X-Match / Spark, Simbad

- Participation (Anais, Vincent, André) to the first GROBID Camp hosted by ResearchGate in Berlin (GROBID is used in DJIN2, see R&D 2016)

- Posters (ADASS 2017, JDEV 2017), Talks at LISA VIII, invitations to present the work (First ASTERICS-OBELICS workshop), collaborations, etc.

Future investigation plans 2018-2021

- Interfaces and interactions
  - High resolution (at least 4K) screens will be common
  - Merging between smartphone / tablet / laptop / desktop Operating Systems probably done during the period
  - Frontier between standalone apps / Web apps getting thinner
  - Other kinds of interactions are emerging (voice, gesture)
  - Continuous R&D in this frame is crucial

Future investigation plans 2018-2021 (2)

- Big Data
  - Continuous R&D effort to provide an added value on the access mechanisms to the data (organisation of the data, metadata, technologies)

  - Providing the tools to access and explore all the CDS data (and external data in the context of the interoperability)
    - Knowledge databases, machine learning, deep learning, log mining etc. to help us
    - Crucial

Future investigation plans 2018-2021 (3)

- Immersive 3D-Visualization, in standby after a few tests in the past years but ready to continue when mature
- Clouds (=> on the rail with Amazon AWS credits)
- Social networks / design & communication
  - CDS logos, flyers, …, video tutorials: done during 2015/17
  - In 2018 we will focus on the social networks (involving deeply the users ?)
- Connected objects
  - A large variety of devices around 2020, probably seen as gadgets in astronomy but many use cases (alerts, news, etc.) which are not yet well identified

□ Conclusion

- From the system & hardware to the natural language processing

- The R&D activity is a way to update and improve the services... and the skills of the staff
  - New technologies
  - Presentation of the work at the end of the internships

- It is not only technical..., it is also a “human” activity making people work together

- Work together means “all the people” participating to the administrative & hardware parts, the presentation of the services and professions – their kindness with the students, a great experience for all of them