

Extraction de données Obscore dans l'ingestion de nouveaux catalogues



Nathan Barloy, Gilles Landais (CDS)



1. Contexte

Indexation des catalogues VizieR avec une base des données associées

The screenshot displays the VizieR web interface. At the top is a navigation bar with logos for CDS, PORTAL, SIMBAD, VizieR, ALADIN, XIMATCH, OTHERS, and HELP. The main heading is "Search associated data among the VizieR catalogues". Below this, a text box explains that the page provides access to VizieR associated data (images, spectra, timeseries, SED) from publications, supervised by the CDS documentalists. A "VO compatibility" section states that the meta-data and search engine are built according to the VO framework (SIA, SSA, ObsTAP) and can be queried by VO softwares, using Saada engines and the ObsCore model.

The search interface includes a "Simple search" tab and an "ObsTAP Query" tab. The search criteria are as follows:

- Search by position: [] radius 1 deg
- Search by spectral band: min 600 max 800 nm
- Search by time data: start [] stop [] (MJD)
- Search by catalog: [] Identifier: []
- Radio buttons: ☐ Spectrum / Time series ☒ Image
- Dropdown: 500 entries max
- Buttons: Search, Tar

On the right, a sky map shows the search area with a field of view (FoV) of 174.67".

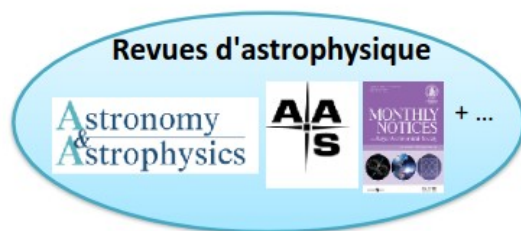
Below the search form, a table displays the results. The table has columns for Preview, Target, Data collection, Ra, Dec, Band min (nm), Band max (nm), Begin time (MJD), End time (MJD), and Facility. The first three rows of data are shown, all for target M31 with data collection J/A+A/526/A155.

Preview	Target	Data collection	Ra	Dec	Band min (nm)	Band max (nm)	Begin time (MJD)	End time (MJD)	Facility
	M31	J/A+A/526/A155	10.645	41.233	356.180	896.150	52,553.000	52,553.001	2.5m
	M31	J/A+A/526/A155	10.645	41.233	356.180	896.150	52,553.000	52,553.001	2.5m
	M31	J/A+A/526/A155	10.645	41.233	356.180	896.150	52,553.000	52,553.001	2.5m

1. Contexte

– Open Data pour l'Astronomie (CDS-ODAS)

Un renouvellement majeur du service de données associées CDS VizierR.



<https://cdsarc.cds.unistra.fr/assocdata/>

<https://vizier.cds.unistra.fr/>



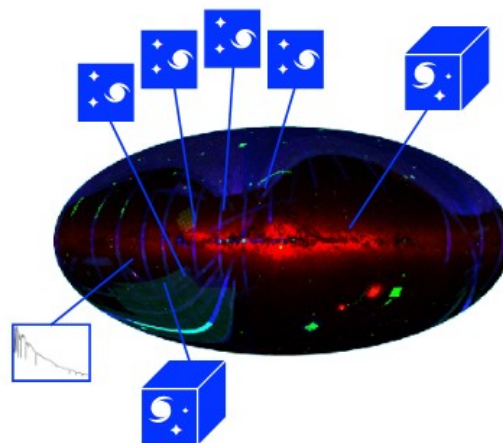
1. Renouvellement du workflow pour le dépôt des données associées aux publications de revues à comité de lecture: images, spectres/séries temporelles, cubes de données

2. Renouvellement du mécanisme d'extraction des métadonnées des cubes / images / séries temporelles au modèle de données standard IVOA.

Études de faisabilité sur différents systèmes d'extraction de métadonnées et d'ingestion des données. (y compris Dataverse)

3. Intégration avec les interfaces de service VizieR rendant les données FAIR via des protocoles d'accès standardisés.

4. Documentation et diffusion à la communauté astronomique - mise en avant des nouvelles capacités de publication des données pour soutenir la science ouverte.

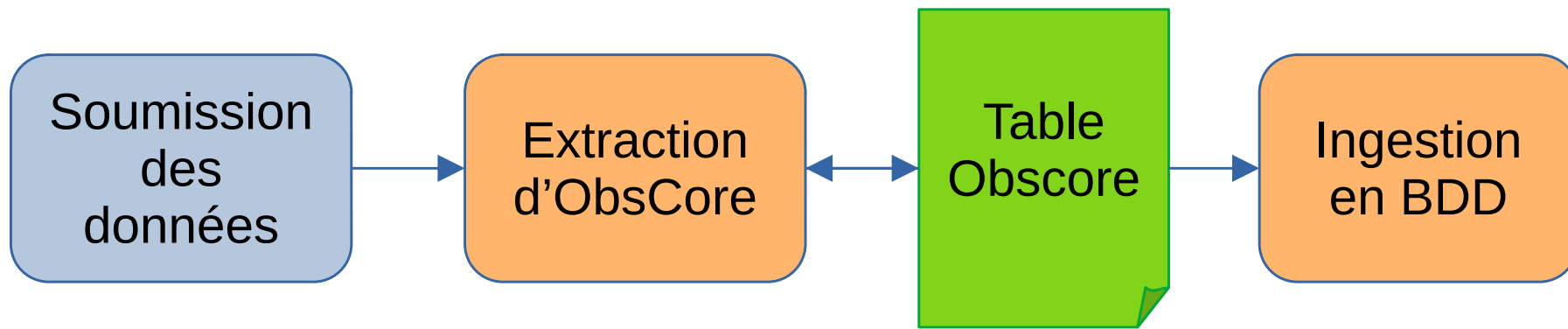


1. Contexte

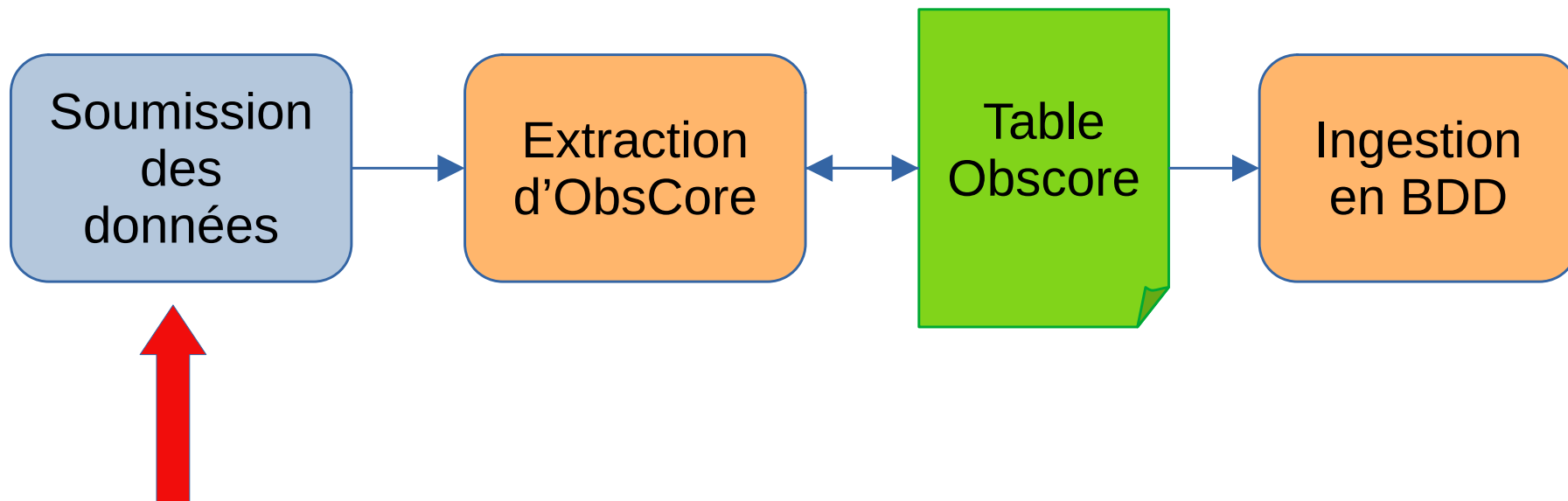
- Nos outils d'ingestion de création et d'ingestion des données ObsCore sont trop liés à l'ancien workflow, il en faut donc de nouveaux.
- Autant en profiter pour créer un outils qui pourra être partagé et distribué.
- Le coeur du nouveau workflow est une bibliothèque python, qui est simple d'utilisation et pourra être facilement distribuée.



2. Aperçu général



3. La soumission des données



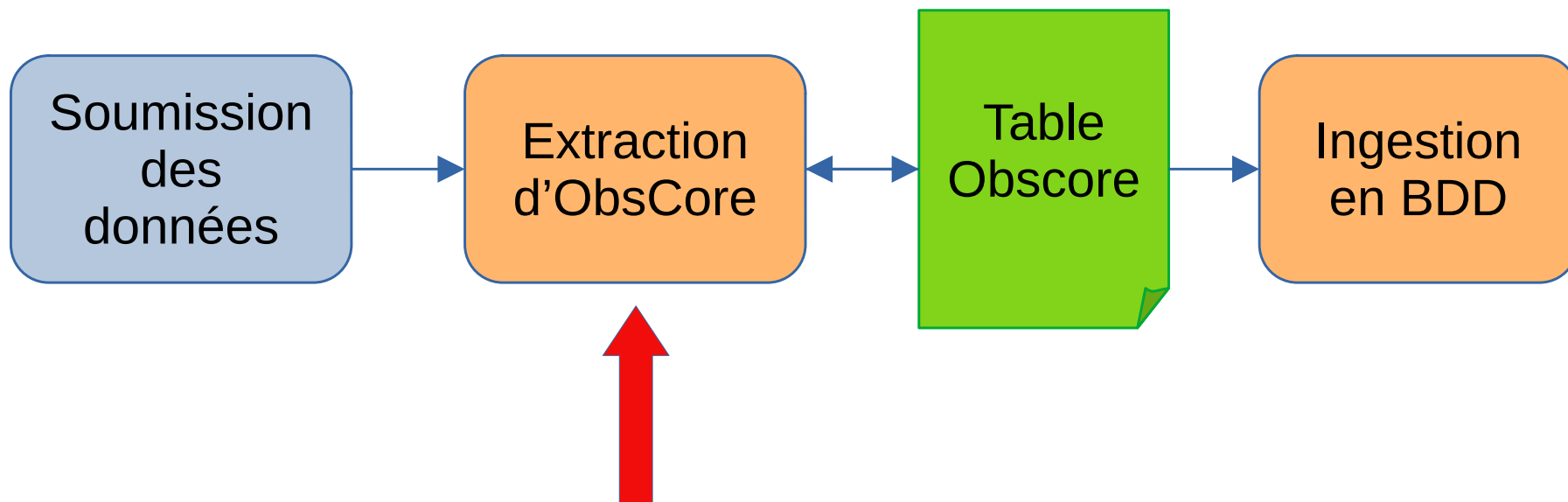
3. La soumission des données

C'est la partie qui n'est pas encore faite du projet. Il y a encore quelques questions sur les choix de design. (Invenio ou Dataverse ?)

Actuellement, il y a un peu moins de la moitié soumissions qui sont faites via la page web, le reste est directement envoyé à nos documentalistes. Il y a donc 2 outils de soumission à créer :

- 1) Une page web publique, pour que les auteurs puisse directement soumettre leurs données. Ils peuvent vérifier directement si les données ObsCore sont correctes.
- 2) Un outils interne pour les documentalistes, simple d'utilisation.

4. Génération des données ObsCore



4. Génération des données ObsCore

Le coeur de cette bibliothèque python est de fournir des fonctions permettant d'obtenir une table ObsCore à partir des fichiers FITS du catalogue.

Il fonctionne en 3 parties :

- 1) Extraction automatique des données depuis les fichiers fits
- 2) Curation des données avec des services externes
- 3) Ajout manuel de données par l'utilisateur

4.1. Extraction des données ObsCore

1) Extraction automatique des données depuis les fichiers fits

- On utilise exclusivement les headers des fichiers FITS
- On se base sur la bibliothèque astropy
- On récupère une partie des données simplement en analysant des keywords FITS classiques.
- Le WCS est résolu avec Astropy, en faisant quelques adaptations si besoin.



4.1. Extraction de données ObsCore

Quelle flexibilité se donner sur les FITS reçus ?

On aimerait bien tout accepter mais...

Des utilisations non standards
arrivent régulièrement

CTYPE1 = 'LINEAR'

On peut ajouter des règles maison
pour ça

Parfois, on peut supposer des
informations mais on ne devrait pas

CTYPE1 = 'WAVE'
mais pas de CUNIT1

Par défaut, l'unité des longueurs
d'onde est le mètre

En général, on essaye d'accepter les règles non standards tant qu'elle
n'entrent pas en contradiction avec des standards déjà existantes.

4.2. Curation des données

2) Curation des données

- On peut aller récupérer des données manquantes dans Simbad.
Typiquement, si on a le nom de l'objet, mais pas ses coordonnées.

- Si les données des bandes sont manquantes, on pourrait les récupérer avec les nom des observatoires des instruments et des filtres.

Mais comment rechercher ça correctement ?



IRS filters:

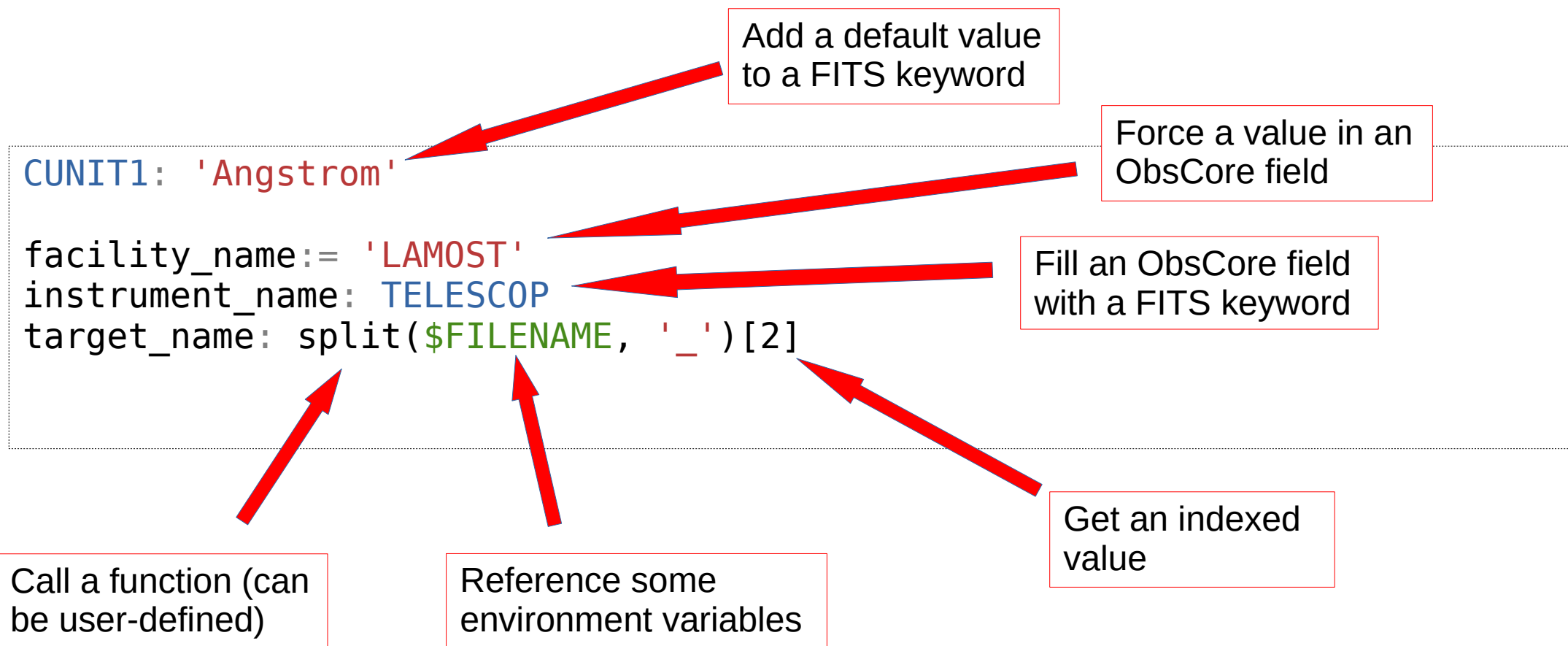
Filter ID	λ_{ref}	λ_{mean}	λ_{eff}	λ_{min}	λ_{max}	W_{eff}	ZP_V	ZP_{λ}	Obs. Facility	Instrument	Description
IRS/MOS.band408	4074.78	4075.25	4070.38	3979.53	4220.00	100.94	4010.04	7.24e-9	IRS	MOS	IRS-P3 MOS band408 RSR
IRS/MOS.band443	4429.16	4429.55	4429.38	4333.52	4509.48	108.11	4369.58	6.68e-9	IRS	MOS	IRS-P3 MOS band443 RSR
IRS/MOS.band485	4837.20	4837.55	4833.11	4745.51	4909.89	105.65	3470.45	4.45e-9	IRS	MOS	IRS-P3 MOS band485 RSR
IRS/MOS.band520	5197.30	5197.61	5197.04	5110.21	5271.82	103.57	3793.18	4.21e-9	IRS	MOS	IRS-P3 MOS band520 RSR
IRS/MOS.band570	5694.41	5694.68	5694.17	5613.92	5768.27	103.59	3484.89	3.22e-9	IRS	MOS	IRS-P3 MOS band570 RSR
IRS/MOS.band615	6148.07	6148.33	6147.82	6072.10	6219.89	107.31	3210.30	2.55e-9	IRS	MOS	IRS-P3 MOS band615 RSR
IRS/MOS.band650	6496.60	6496.85	6495.15	6422.05	6577.34	108.54	2929.60	2.08e-9	IRS	MOS	IRS-P3 MOS band650 RSR

4.3. Le mapping

3) Ajout manuel de données par l'utilisateur

- Motivé par le fait qu'on ne peut pas directement modifier les FITS qu'on nous envoie. De plus, certains catalogues ont un même format.
- Ce fichier de mapping permet de donner des informations complémentaires, en faisant une simple correspondance clé/valeur.
- Ce fichier peut non seulement être donné lors de l'étape d'extraction, mais il peut aussi être appliqué directement sur une table ObsCore déjà existante.

4.3. Le mapping



4.4. Utilisation

Utilisation extrêmement simple dans python

```
from assocdata import extract, apply_mapping, load_table

catalog_directory = '~/Documents/catalog1.fits'
response = extract(catalog_directory)
response.curate()
obscore_table = response.to_table()

# With a mapping
response = extract(catalog_directory, './path/to/mapping_file.map')

# Save and load a table
response.write('obscore.csv')
table = load_table('obscore.csv')

# We can also apply a mapping afterwards
apply_mapping(obscore_table, './path/to/mapping_file.map')
```

4.5. Utilisation de LLM (Diana Ismail)

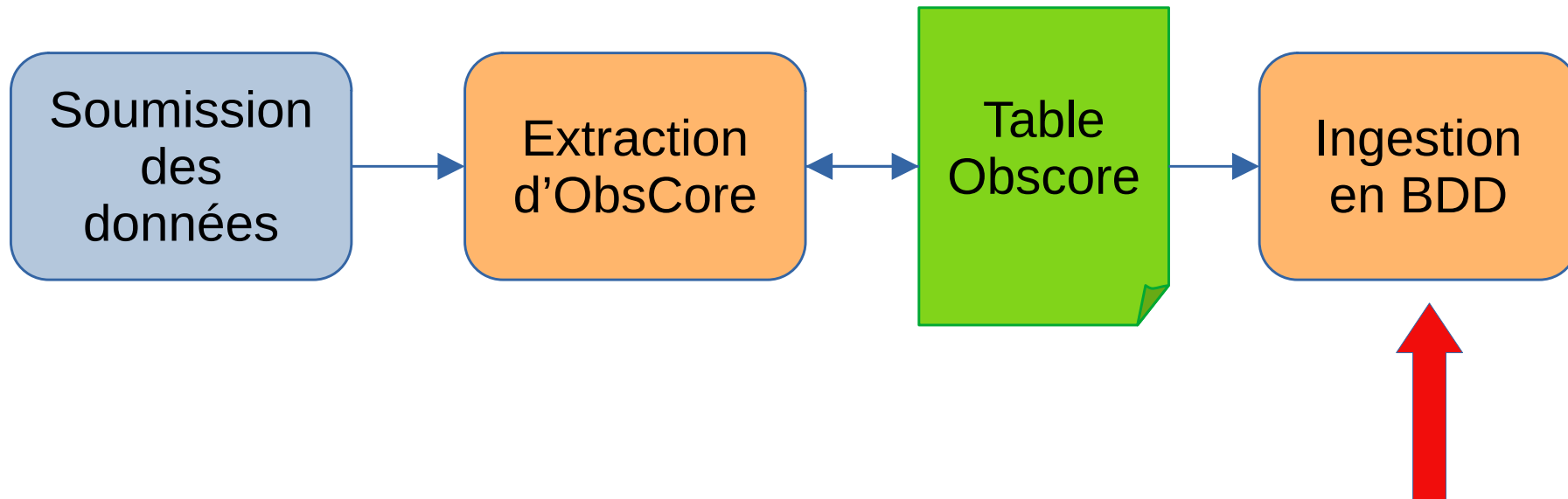


ChatGPT



- Les tests n'ont pas été très concluants pour extraire des informations depuis les headers FITS, mais l'utilisation du fichier readme est plutôt prometteuse.
- Après plusieurs benchmark, Gemini semble donner les meilleurs résultats à moindre coûts. (Très variable dans le temps)
- Surprenamment, les résultats sont meilleurs si on lui demande de réfléchir modérément.
- Cet outils peut assez bien compléter la bibliothèque python.
- Cela pose tout de même quelques questions de droit. Peut-on envoyer des données pas encore publier dans des LLM privés ?

5. Ingestion en Base de Données



5. Ingestion en Base de Données

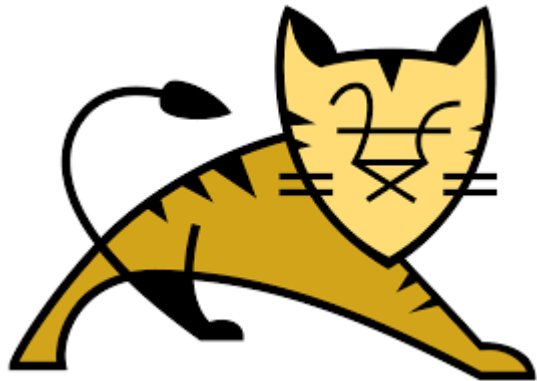
Le format de la base de données ne change pratiquement pas de l'ancien : il s'agit d'une base PostgreSQL avec l'extension PgSphere.



Possible changement : utiliser un nouvel index sur les positions développé en interne à l'occasion d'un stage. Cela permettra de le tester.

5. Ingestion en Base de Données

Comme avant, l'accès à cette base de données se fait via un serveur TAP. Ce serveur est créé avec notre outils Vollt, et est installé via apache tomcat.



6. Ouverture de la bibliothèque

A terme, la bibliothèque d'extraction de table obscure devrait être disponible en open source.

Le développement s'est fait en pensant à l'utilisation faite en interne au CDS, donc il reste un certain nombre de points à adapter.

On pourrait aussi ajouter quelques petits outils supplémentaires, comme la génération de miniature par exemple.