CENTRE DE DONNÉES
ASTRONOMIQUES DE STRASBOURG

Renouvellement du service d'ingestion de catalogue VizieR

G.Landais , I. Brossard, N. Barloy, F-X. Pineau, A. Vanhule

Le service d'ingestion VizieR

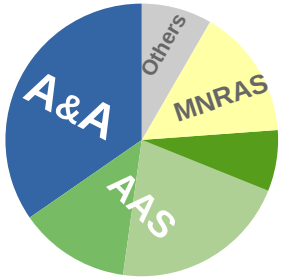
Le processus d'ingestion VizieR

- Mise en ligne des données (principalement tabulaires) issus d'articles, de missions terrestres ou spatiales
- Crée par F. Ochsenbein en **1995**
- Service certifié par le CTS (2018)

Statistiques

- 25,839 catalogues
- 62,694 tables
- **1,059,497 colonnes**
- 8,312,847 spectres
- 51,431 images
- ~ **1,200 nouveaux catalogues/an**

Origine des catalogues VizieR



La FAIR-isation des données

Données traitées

- Les données sont mise en ligne par une équipe de documantliste
- Des flux de données propres a l'origine des données (ex : Grand catalogues, A&A, AAS, …)
  MAIS tous utilisent le même corpus et formatde méta-données
- Des données majoritairement statiques (versioning et mise à jour rare)

Distribution

- L'entité est le catalogue qui est un "dataset" dont tous les éléments sont documentés
- Les données sont accessibles via une base de données unique et via différent formats et protocoles
- Distribution des métadonnées dans l'IVOA, ADS, EOSC, Recherche DataGouv.

Objectif de la migration

Essentiellement technologique

- Un code modulable, plus simple à faire évoluer
- Utiliser des technologies plus récentes : C → Python

Conservative

- Doit pouvoir remettre en ligne les données historique (>98%)
- Conservation des méta-données existantes
- Doit s’insérer dans l’écosystème actuel (VizieR, CDS, IVOA ...)
- Conserver les habitudes des personnes en charge de l’ingestion

Objectif de la migration

Une migration du code qui s'insère dans la maintenance évolutive de VizieR

- Migration de SGBD Sybase → PostgreSQL (G.Landais, 2008)
- Recherche textuelle → ElasticSearch (Y.Bish, G.Landais, 2019)
- Indexation globale : Qbox → MOC (mocset, moc-cli, F-X. Pineau, 2023)
- Migration du système de gestion des grands catalogues : → AT2S (F-X. Pineau, 2024)
- Migration du système de gestion du traitement des données Spectres et images (N.Barloy)

Charge de travail

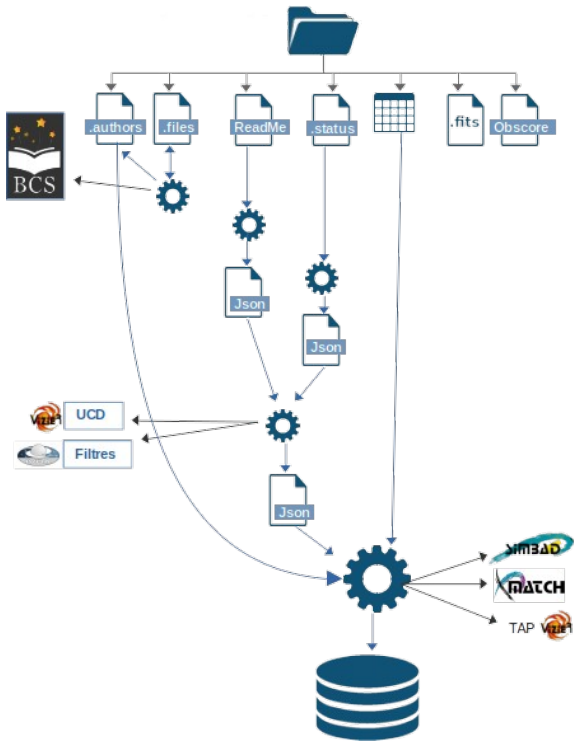
- Début des travaux : 2021, 2022
- Développeurs :
  - Permanent : G.Landais (20%), F-X.Pineau
  - contractuels : A.Vanhule (05/2022 → 11/2023), I. Brossard (05/2024 → ...)

Principe du traitement de mise en ligne

Le processus d'ingestion

- Programme en ligne de commandes (Linux) à l'usage des documentalistes du CDS
- Système **semi automatisé**
  - Extraction «**directe**» de méta-données:
    - Fichier ReadMe (générés par les auteurs ou les documentalistes)
    - Fichier de configuration (généré par les documentalistes)
    - Ressource externes (ex : BCS pour ORCID ou DOI article)
  - Extraction « **indirecte** » automatique de méta-données partir de textes issues (règles induites)

    Détection de filtres, UCD, association de colonnes
    Ex : erreur, ou alors rassembler positions et mouvement propres
  - Opération sur les tables effectués , paramétrable dans le fichier de configuration



Principe du traitement de mise en ligne

Les grandes étapes / challenges

- Le choix des technologies
- Extraction des méta-data et construction d'une structure de données réutilisable
- Interface utilisateur pour le mapping UCD1+
- Compréhension des règles induites (filtres, etc)
- Calculs sur les mesures , les unités et les Systèmes de coordonnées
  eg : changement de système de coordonnées (mouvement propres), calcul de positions à partir d'offsets et de pixels, etc.
- Opération de crossmatch : jointures, ajout de positions à partir de Simbad, VizieR

J/MNRAS/320/451/table6 A-G star metallicity (Beers+, 2001) 2001MNRAS.320..451B ReadMe+ftp
Post annotation
Derived quantities - EC cool stars *(206 rows)*

start AladinLite plot the output query using TAP/SQL

Full	EC	Class	u_	VMAG mag	VMAG2 mag	Dist pc	Dist2 pc	[Fe/H]K3 Sun	u_	e_	Beers	Simbad	_RABcomp deg	_DEBcomp deg	_RA.icrs deg	_DE.icrs deg
1	00097-3243	SG		3.27		4192		-2.44		0.18	Beers	Simbad	002.42	-32.72	003.0572	-32.4385
2	00230-2340	TO		4.64		766		-0.44		0.27	Beers	Simbad	005.75	-23.67	006.3779	-23.3898
3	00316-2543	TO		4.23		1332		-0.55		0.15	Beers	Simbad	007.90	-25.72	008.5216	-25.4411
4	00341-7725	TO		4.46		593		-0.98		0.15	Beers	Simbad	008.52	-77.42	008.9780	-77.1414
5	00347-7711	TO		4.52		927		-1.11		0.15	Beers	Simbad	008.67	-77.18	009.1284	-76.9082
6	00448-6351	SG		3.60		1211		-1.35		0.29	Beers	Simbad	011.20	-63.85	011.7286	-63.5771

Les metadonnées de bases

Le fichier ReadMe

- Le fichier ReadMe (format ASCII) décrit les paquets (paquet SIP).
- Un format lisible (human readable)
- Contient les informations de bases, nécessaires à la **préservation** sur le long terme:
  auteur, abstract, description des colonnes, etc.

```
I/320              SPM 4.0 Catalog              (Girard+, 2011)
================================================================================
Yale/San Juan Southern Proper Motion Catalog 4 (SPM4)
    Girard T.M., van Altena W.F., Zacharias N., Vieira K., Casetti-Dinescu D.I.,
    Castillo D., Herrera D., Lee Y.S., Beers T.C., Monet D.G., Lopez C.E.
   <Astron. J. 142, 15 (2011)>
   =2011AJ....142...15G
   =2011yCat.1320....0P
================================================================================
ADC_Keywords: Positional data ; Proper motions ; Photometry, photographic ;
              Photometry, CCD
Keywords: Astrometry - Catalogs

Description:
    The SPM4 Catalog contains absolute proper motions, celestial
    coordinates, and B,V photometry for 103,319,647 stars and galaxies
    between the south celestial pole and -20 degrees declination. The
    catalog is roughly complete to V=17.5. It is based on photographic
    and CCD observations taken with the Yale Southern Observatory's
    double-astrograph at Cesco Observatory in El Leoncito, Argentina.
    The first-epoch survey, taken from 1965 to 1979, was entirely
    photographic. The second-epoch survey is approximately 1/3
    photographic (taken from 1988 to 1998) and 2/3 CCD-based (taken
    from 2004 through 2008).

    Full details about the creation of the SPM4.0 catalog can be found
    in the paper, and also in the document "spm4_doc.txt" file which
    describes the original files, accessible from
    http://www.astro.yale.edu/astrom/spm4cat/

File Summary:
--------------------------------------------------------------------------------
 FileName   Lrecl  Records   Explanations
--------------------------------------------------------------------------------
ReadMe         80        .   This file
spm4.sam      165     1000   Sample of SPM4 output (out of 103319647 stars)
spm4_doc.txt   80      671   Documentation of the SPM Catalog
--------------------------------------------------------------------------------

See also:
    http://www.astro.yale.edu/astrom/spm4cat/ : SPM Catalog 4 Home Page
    I/289 : UCAC2 Catalogue (Zacharias+ 2004)
    I/322 : UCAC4 Catalogue (Zacharias+ 2012)
   II/246 : 2MASS All-Sky Catalog of Point Sources (Cutri+ 2003)
  VII/233 : The 2MASS Extended sources (IPAC/UMass, 2003-2006)

Byte-by-byte Description of file: spm4.sam
--------------------------------------------------------------------------------
   Bytes Format Units   Label     Explanations
--------------------------------------------------------------------------------
   1- 10 I10   ---    SPMID      Unique identifier (4)
  12- 22 F11.7 deg    RAdeg      Right ascension (J2000, epoch=2000.0)
  24- 34 F11.7 deg    DEdeg      Declination     (J2000, epoch=2000.0)
  36- 40 F5.1  mas    e_RAdeg    Error in right ascension at Epoch
  42- 46 F5.1  mas    e_DEdeg    Error in declination at Epoch
  48- 55 F8.2  as/yr  pmRA       Absolute proper motion in RA (*cos(DE))
  57- 64 F8.2  as/yr  pmDE       Absolute proper motion in Declination
  66- 71 F6.2  as/yr  e_pmRA     Error in pmRA at Epoch
  73- 78 F6.2  as/yr  e_pmDE     Error in pmDE at Epoch
  80- 86 F7.2  yr     Epoch      [1965/2007] weighted mean epoch of observation
  88- 94 F7.2  yr     Ep1        [1965/1982] weighted mean 1st-epoch
  96-102 F7.2  yr     Ep2        [1987/2007] weighted mean 2nd-epoch (2)
 104-105 I2    ---    N1         [1/20] Number of 1st-epoch plates used
 107-108 I2    ---    N2         [0/14] Number of 2nd-epoch plates used
 110-110 I1    ---    Nc         [0/9]  Number of 2nd-epoch CCD frames used (3)
 112-116 F5.2  mag    Bmag       ? Johnson B magnitude estimation (1)
 118-122 F5.2  mag    Vmag       ? Johnson V magnitude estimation (1)
 124-124 I1    ---    ib         [1/3] source flag for B mag (1)
 126-126 I1    ---    iv         [1/3] source flag for V mag (1)
 128-133 F6.3  mag    Jmag       ? 2MASS J magnitude
 135-140 F6.3  mag    Hmag       ? 2MASS H magnitude
 142-147 F6.3  mag    Kmag       ? 2MASS Ks magnitude
 149-149 I1    ---    igal       [0/3] galaxy/extended-source flag (5)
 151-151 I1    ---    icat       [1/7] input catalog source flag (6)
--------------------------------------------------------------------------------

Note (1): ib and iv give the source of the magnitude estimation as:
     1 = compiled input catalog
     2 = 1st-epoch SPM plates
     3 = 2nd-epoch SPM CCD observations
```

Les métadonnées avancées

Le fichier de configuration d'ingestion d'un catalogue VizieR

- Le fichier de configuration (format ~LaTeX) contient les informations "riches" qui permettent d'avoir des données **interopérables**

  Ex: UCD, filtres, définitions de liens, opération de crossmatch, etc

- Utilisation d'un **langage de macro** (basé sur LaTeX)

Exemple de commandes

- Déclaration d'une opération d'UNION

```
\vizMerge{ table[567] }{ plx }{ Parallax and motion fits for objects with high-quality
    (Table 5), low-quality (Table 6) and poor-quality (Table 7) parallaxes }
```

- Déclaration d'une opération de jointure

```
\vizPaste{ tablec2 tablec3 }{ galaxies }{ Model and galaxy parameters }
```

- Déclaration de l'ajout d'une colonne calculée (créant un lien sur une table externe avec le nombre de records)

```
\vizAddCount{ table1 }{ Nf }{table4 2MASS=@{2MASS}}{dbtype=i2 fmt=3d}{}
  {\use{vPop}\ucd{NUMBER}Number of fit results for this star (link to Table 4)}
```

- Assignation d'un UCD

```
\vizUCD{ table6 }{ * Ratio* }{ =phot.flux;arith.ratio;em.radio }
```

- Déclaration de l'ajout de colonnes de position depuis un catalogue VizieR (~jointure sur table3.Seq=table5.ID ...)

```
\vizPosition{ table5 table6 table7 table8 }{}{source=J/ApJS/235/4/table3 Seq=@{ID} Eq=J2000 Ep=2000 precoo=5 }
  { Positions from the 105-month Swift-BAT survey (Oh+, 2018, J/ApJS/235/4) }
```

- Ajout de colonnes "lien" sur Simbad (possible utilisation d'expressionde tests complexes)

```
\vizSimbad{ table1 }{ \ifmatch{VESTA}{${Name}}{}\else{2MASS J${Name$3-}}\fi}
\vizSimbadName{ table8 } { table8.sim }
```

Structurer les méta-données

Transformation des informations au format JSON

- ReadMe
- Ficher de config (LaTeX)
- UCD


```
{
  "expression": [
    {
      "latex": {
        "key": "\\vizSimbad",
        "parameter": [
          {
            "expression": {
              "expression": [  { "string": "table1"  }  ]
            }
          },
          {
            "expression": {
              "expression": [
                {
                  "latex": {
                    "key": "\\ifmatch",
                    "parameter": [
                      {
                        "expression": {
                          "expression": [ { "string": "VESTA" } ]
                        }
                      },
                      {
                        "expression": {
                          "expression": [ { "ref": "${Name}" } ]
                        }
                      }
                    ]
                  }
                } ....
}
```

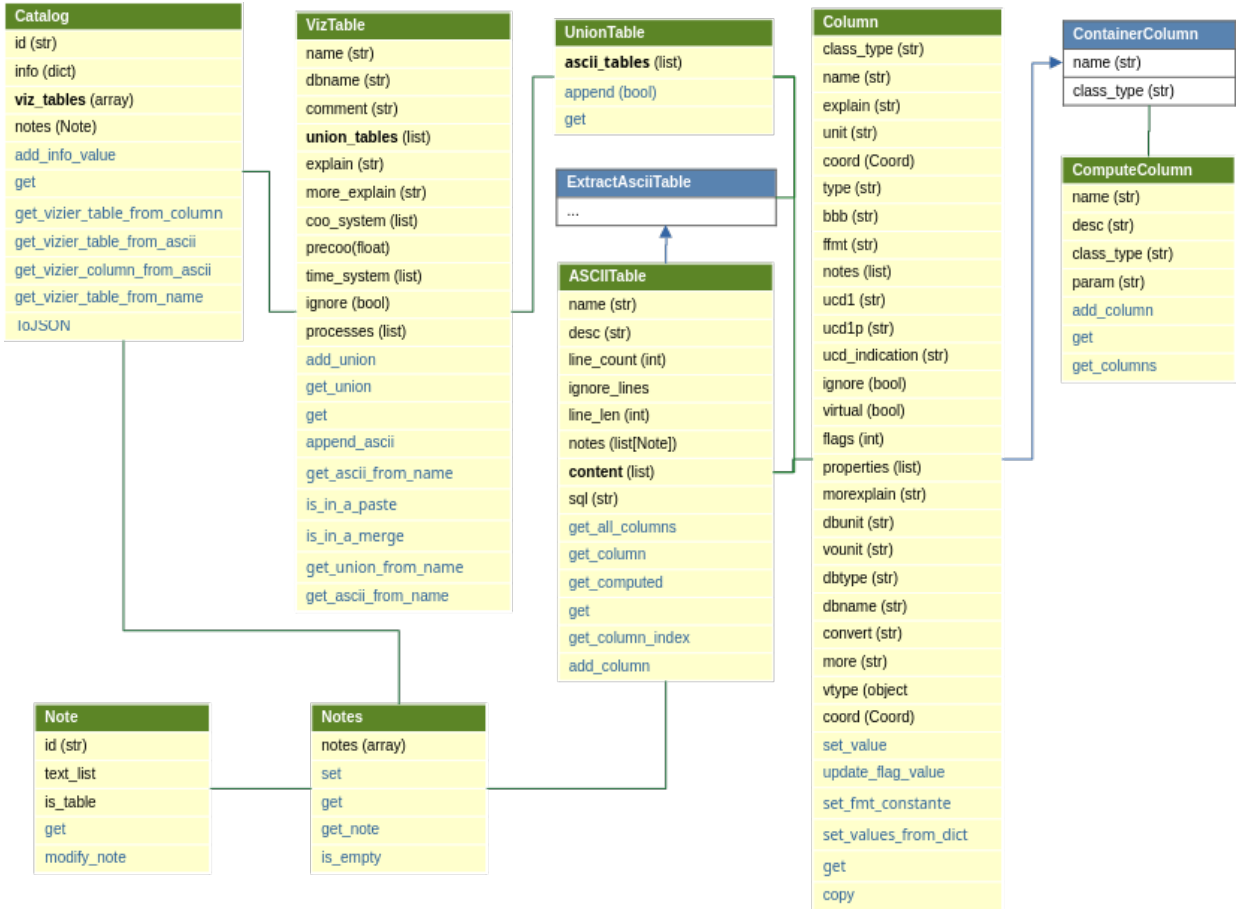

Extrait de la sérialisation JSON du fichier de config


```
ID:                          "J/A+A/267/515"
▶ global_info:               { title: "Very cold C-rich circumstellar envelopes (Omont+ 1993)", year: "1993", authors: "Omont A., Loup C.,
                             Forveille T., te Lintel Hekkert P.,Habing H., Sivagnanam P.", … }
▼ viz_tables:
  ▼ 0:
      name:                  "table1"
      dbname:                "c32670515t1"
      explain:               "Supergiants"
    ▼ union:
      ▼ 0:
        ▼ join:
          ▼ 0:
              name:          "table1.dat"
              desc:          "Supergiants"
              line_count:    19
              line_len:      58
            ▼ columns:
              ▶ 0:           { class_type: "compute", name: "add_recno", param: "", … }
              ▼ 1:
                  class_type: "simple"
                  name:      "No"
                  dbname:    "no"
                  explain:   "Identification number in tablea1 and tablea2"
                  unit:      null
                  dbunit:    null
                  vounit:    null
                  bbb:       "1-3"
                  ffmt:      "A3"
                  fmt:       "%-3s"
                  type:      "VT_CHAR"
                  dbtype:    "VT_CHAR"
                  length:    3
                  flags:     "VO_DISPLAY | VO_SELECT | VO_FK"
                  ucd1p:     "meta.id"
              ▶ 2:           { class_type: "simple", name: "Name", dbname: "name", … }
              ▶ 3:           { class_type: "simple", name: "Sp", dbname: "sp", … }
              ▶ 4:           { class_type: "simple", name: "Region", dbname: "region", … }
              ▶ 5:           { class_type: "simple", name: "LRS", dbname: "lrs", … }
              ▶ 6:           { class_type: "simple", name: "FIRAS", dbname: "firas", … }
              ▶ 7:           { class_type: "simple", name: "n", dbname: "n", … }
  ▶ 1:                       { name: "table2", dbname: "c32670515t2", explain: "Sources with CO(2-1) or HCN(1-0) emission", … }
  ▶ 2:                       { name: "table3", dbname: "c32670515t3", explain: "Non-detected sources", … }
  ▶ 3:                       { name: "table4", dbname: "c32670515t4", explain: "CO/HCN line ratios", … }
  ▶ 4:                       { name: "table5", dbname: "c32670515t5", explain: "Estimates of mass-loss rates", … }
  ▶ 5:                       { name: "master", dbname: "c32670515master", explain: "Master list of very cold C-rich sources,   main sample
                             (table A1) and additional sources (table A2)", … }
▼ global_notes:
```

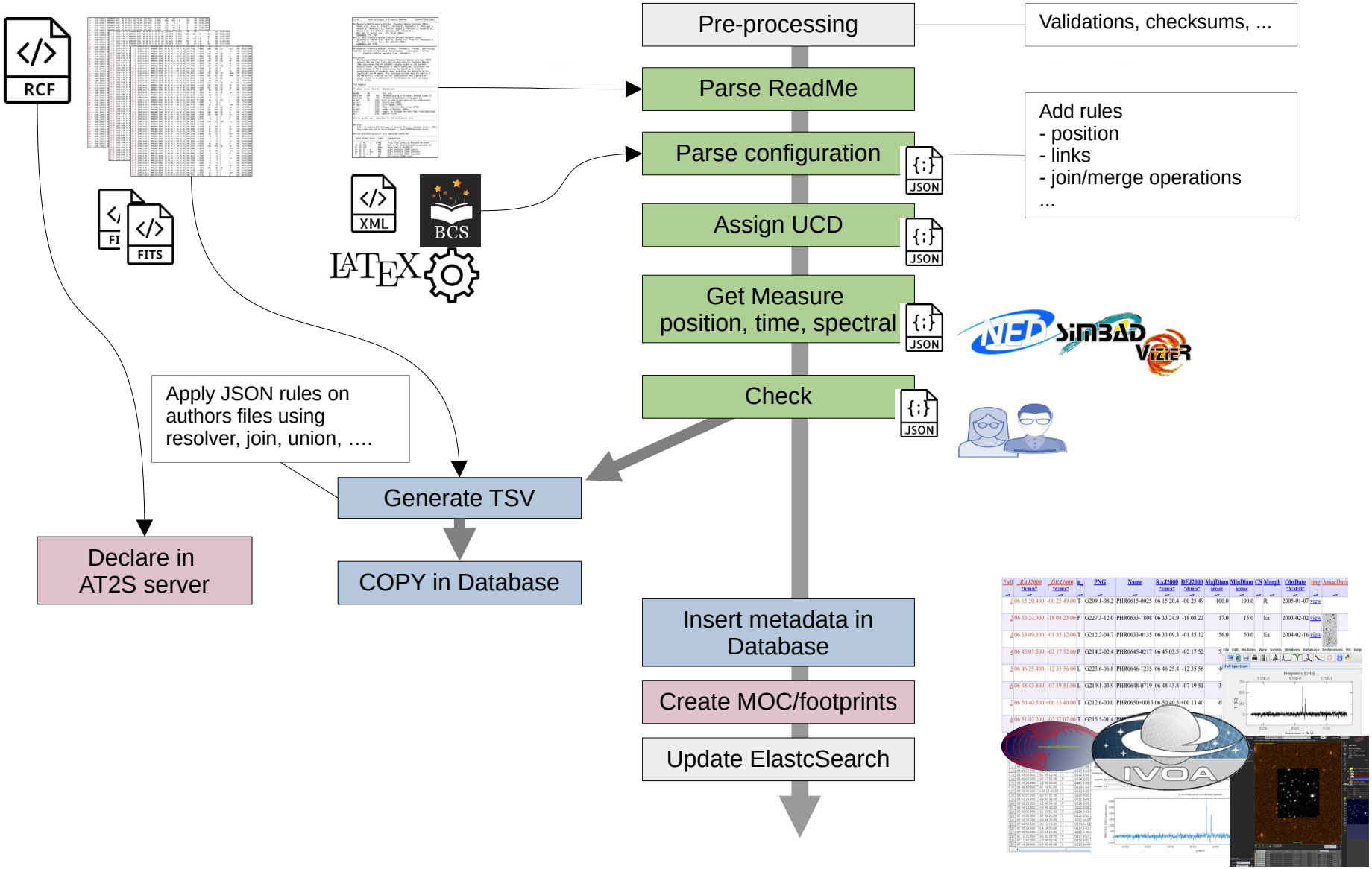

Extrait de la sérialisation JSON de l'ensemble des méta-donneées pour l'ingestion d'un catalogue

Structurer les méta-données

Diagramme de classes, utilisé par les API d'ingestion du catalogue VizieR



Etapes du processus d'ingestion



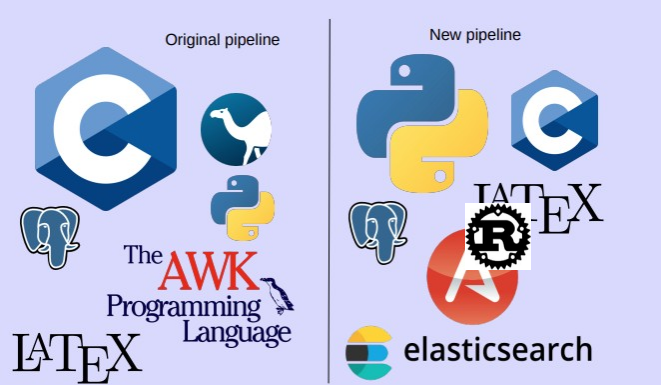
Technologies

Technologies

Languages: C, AWK, Cshell, LaTeX

Database: PostgreSQL

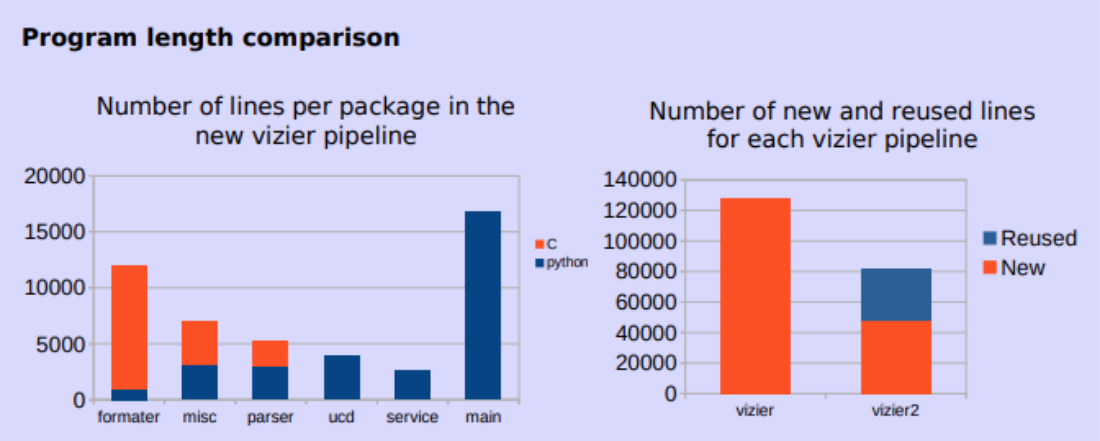
Dépendances: unity, healpix (nasa), erfa



Languages: C/C++, Rust, Python, LaTex, BASH, ANTLR4

Databases: PostgreSQL, ElasticSearch, AT2S

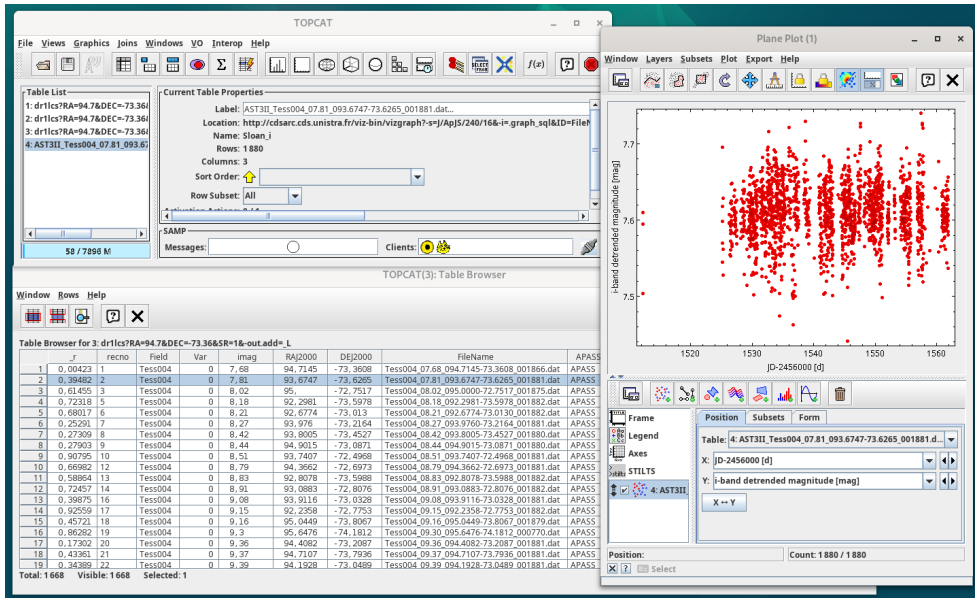
Dépendances: unity, healpix (nasa), erfa, antlr



Graphes par I.Brossard, Poster ADASS, Goerlitz 2025

Quoi de 9 ?

- Il s'agit principalement d'une migration technologique
- Modification sur le stockages des colonnes float (stocke en float/double précision) avec gestion des valeurs NULLES
- Mais quelques nouveautés :
  - Intégration UCD1 → UCD1+ dans le workflow
  - Exploitation des liens en Datalink (cf I.Brossard, IVOA Goerlitz)
  - Indexation globale MOC (en prod)
  - Affiliation et ORCID généralisés (en cours)
  - Ajout des licences (en cours)



Extraction d'une série temporelle VizieR via TOPCAT

Conclusions

- Une migration longue - 4 ans de labeurs
- ~130K lignes de codes, écrites en C dans un « ancien style »
- Manque de documentation :
  De nombreuses règles induites, non documentées
- Des partis -pris historique n'ayant plus de raisons pratiques
- Un code plus optimisé pour l'ingestion des tables

```
(INFO) - database  - Database Connection (psycopg): dbname=vizier user=asu port=5433 host=127.0.0.1
==> initialize database access     [OK]
==> Parse Readme                   [OK]
==> Parse status                   [OK]
==>                                [OK]
(INFO) - database  - Database Connection (psycopg): dbname=vizier user=asu port=5433 host=127.0.0.1
(WARN) - metabuild - \vizFoot not managed yet (\vizFoot{-scale l})
(WARN) - metabuild - undefined vizSet attribute atEnd
(WARN) - metabuild - min dbtype i2t is ignored
==> Build basics metadata          [OK]
==> Analyse UCD                    [OK]
==> Build columns metadata         [OK]
(INFO) - database  - Database Connection (psycopg): dbname=vizier user=asu port=5433 host=127.0.0.1
(WARN) - metabuild - Column u_Epoch doesn't look as a Time column
==> Build measures metadata        [OK]
(ERRO) - metabuild - extract columns from tex \ifregexp{[A-Za-z0-9][A-Za-z0-9]* [A-V][A-Za-z][a-z]}{@{}}   \wSim{V* @{}}{@{}}\else{@{}}\fi: Latex wSim is not defined: abort
==> Build computed metadata        [OK]
==> Build resolver metadata        [OK]
==> last check update              [OK]
==> Check indexes                  [OK]
==> check UCD                      [OK]

--------------------------------------------------------------------------------
REPORT ReadMe/status analysis
  - Table vsx (paste) vsx.dat,/home/cats/tmp/gl/vsx_n_OID.match
    Position: FK5 , equinox=2000 (RAJ2000,DEJ2000,,,,)
    Time: Epoch system=HELIOCENTER_default,rep=JD
--------------------------------------------------------------------------------
(INFO) - database  - Database Connection (psycopg): dbname=vizier user=asu port=5433 host=127.0.0.1
(INFO) - database  - Database Connection (psycopg): dbname=vizier user=cds port=5433 host=127.0.0.1
==> remove catalogue               [OK]
(INFO) a2csv: file /home/cats/B/vsx/refs.dat is gziped

[■■■■■■■■■■■■■■■■■■■■.....................................] (3433601/10300781)
```