VizieR 2
Ingestion des données dans VizieR

- Astronomes: P.Ocvirk, C.Bot, H.Arrab
- Ingénieurs: G.Landais, T.Boch, FX-Pinneau
- Documentalistes: P.Vannier, E.Perret, M.Brouty, T.Pouvreau

CENTRE DE DONNÉES
ASTRONOMIQUES DE STRASBOURG

VizieR 2 ?

De quoi il s'agit ?

- Refonte de l'**ingestion des données** table dans VizieR
- Il s'agit principalement d'une migration technique

Qu'est ce que VizieR2 n'est pas?

- Il ne s'agit **PAS** de l'accès aux données (service, page web...)
- Ce n'est **PAS** dans le but d'améliorer le travail de documentation
- Il ne s'agit **PAS** de révolutionner les méthodes d'ingestion

But

- **framework plus intégré au service du CDS et plus modulable!**

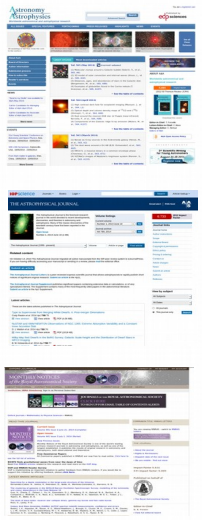
What is VizieR ?

Vizier gives a unified access to a very large collection of astronomical catalogues

- Provides a **free** access to **public** catalogues
- Long term **preservation (DSA)**

The content origin

- **Tables** from papers published in the major **astronomical journals**
- **Reference catalogues & surveys** e.g. Gaia, SDSS, 2MASS, UCAC, WISE
- **Logs of observations** and incremental datasets updated periodically
- **Associated data:** spectra, images, time-series



VizieR in numbers

~17,000 catalogues,
~37,000 tables

Associated data:

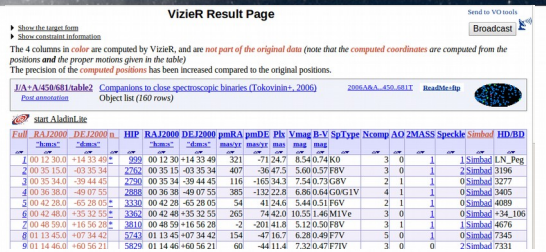
~500 cats. with spectra
~200 cats. with images
~1,200 cats. with time-series

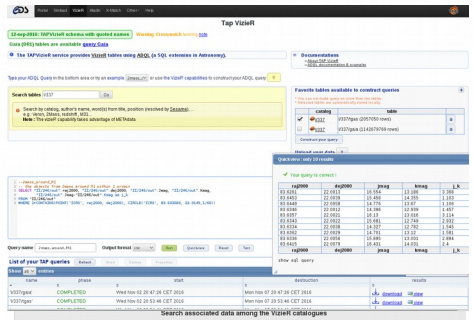
.. But VizieR is also a team

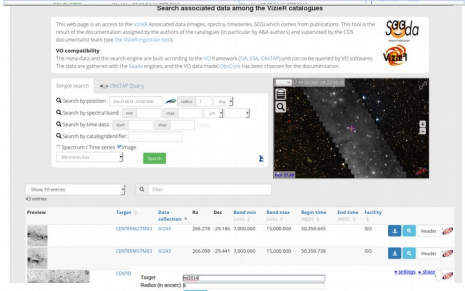
a close collaboration between astronomers, documentalists and engineers

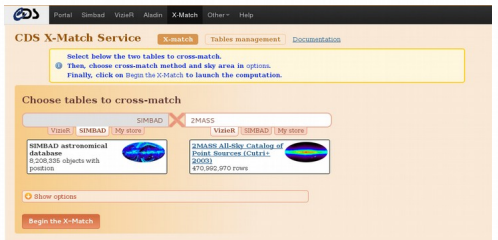
Accéder aux données VizieR

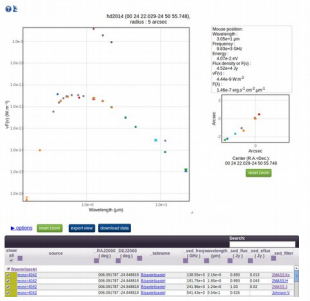
- Accès aux tables par des interfaces dédiées
- Interface VO pour acceder aux tables
  - Format VOTable
  - TAP
  - images (SIA)
  - spectres(SSA)
- Exploitation des données par des services
  - SED VizieR
  - Xmatch
  - (Aladin) MOCS/HIPS





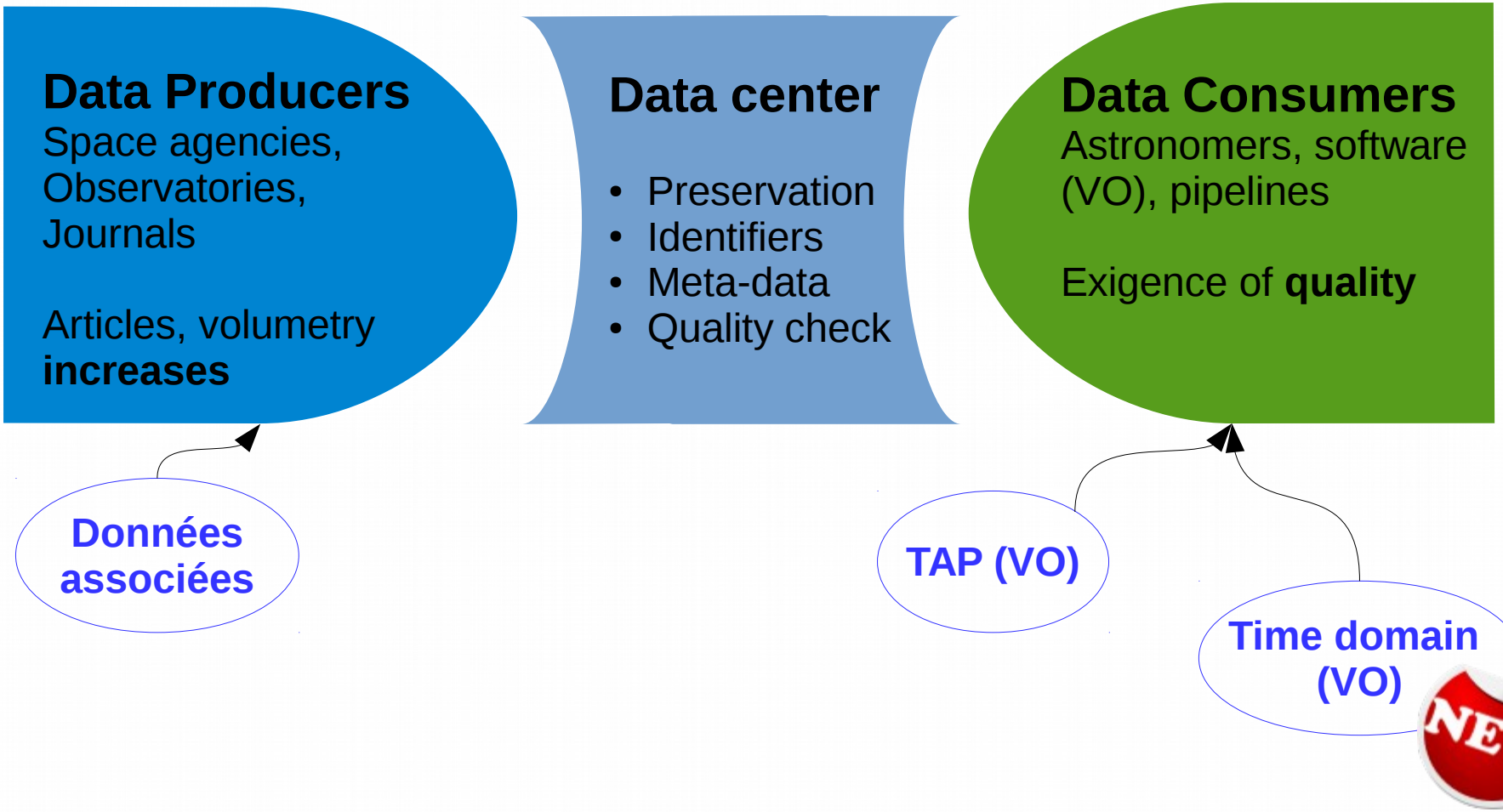






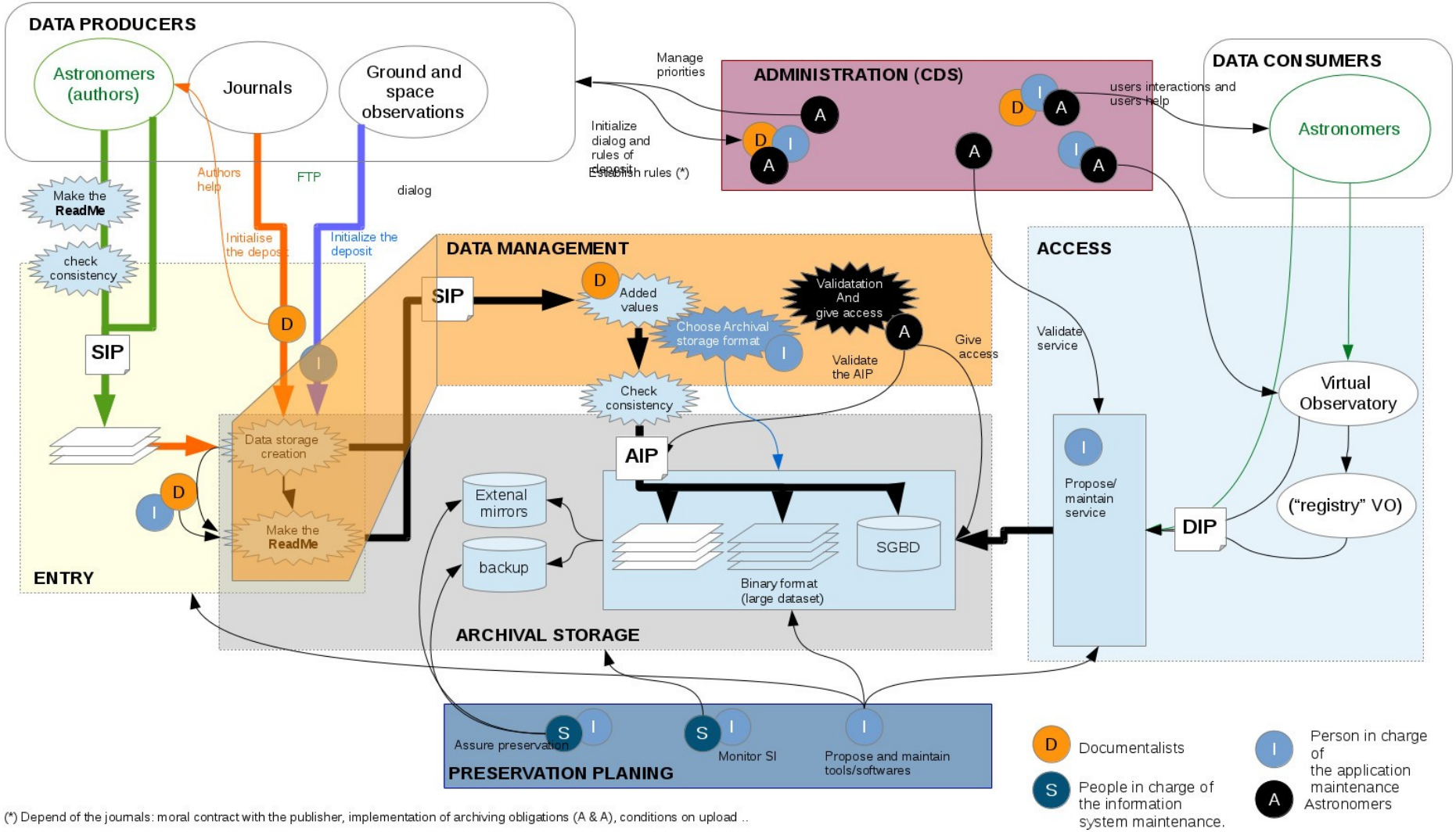
Open Data : impacts on Data Centers

A challenge for Data Centers to deal with increasing volume in input and to ensure quality in output



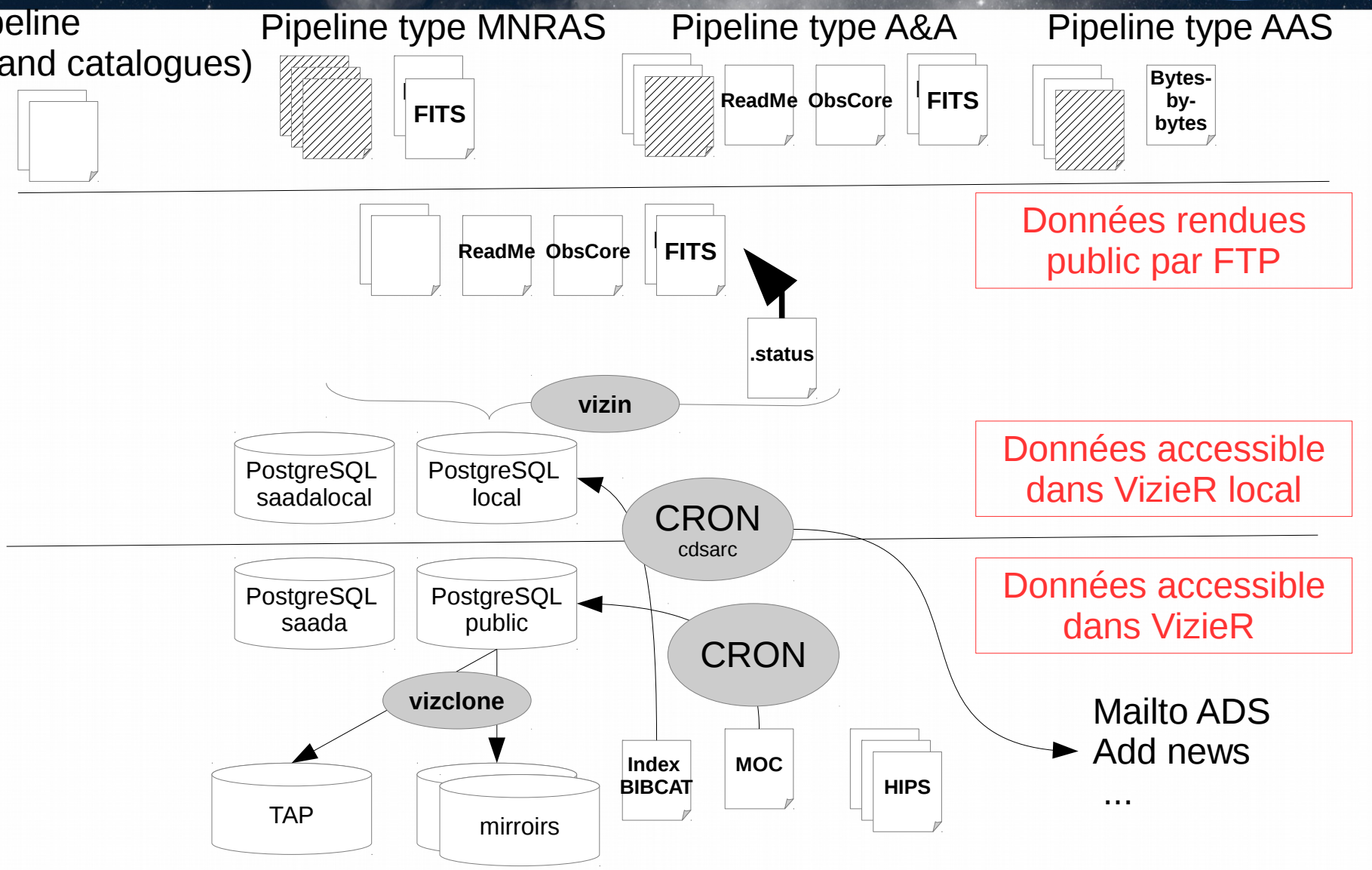
Ingestion VizieR

Etapes d'ingestions dans un processus de péservation des données



(*) Depend of the journals: moral contract with the publisher, implementation of archiving obligations (A & A), conditions on upload ..

Pipelines VizieR



Données en entrées (tables)

Les tables

les données originales sont préservées dans un répertoire “ori” puis standardisées au format ASCII CDS

Note: la précision des données dans le format standardisé peut être modifiée

Métadonnées en entrées (pour les tables)

- **Métadonnées de base** :
  le fichier **ReadMe** - format ASCII lisible
  - Description catalogue/table/colonne (unit, format)
  - Abstract, auteur, bibcode, keywords

  Mais aussi des métadonnées “avancées”
  - Filtres, système de coordonées..
- **Métadonnées “riches”** : le fichier de configuration .status (langage de commandes **LaTeX**)
  - Opérations sur les tables : Jointure/concat, ajout de position, Index, PK
  - Enrichissement des données : UCD, Links ext/int, Wrapper plots, format grand cat
- Autres fichiers utiles à la compréhension des données : mail, plots, données associées

```
J/A+A/424/545       Optically faint obscured quasars         (Padovani+, 2004)
================================================================================
Discovery of optically faint obscured quasars with Virtual Observatory tools.
    Padovani P., Allen M.G., Rosati P., Walton N.A.
   <Astron. Astrophys., 424, 545-559 (2004)>
   =2004A&A...424..545P
================================================================================
ADC_Keywords: QSOs ; Active gal. nuclei ; X-ray sources
Keywords: astronomical data bases: miscellaneous - methods: statistical -
          galaxies: quasars: general - X-rays: galaxies

Abstract:
    We use Virtual Observatory (VO) tools to identify optically faint,
    obscured (i.e., type 2) active galactic nuclei (AGN) in the two Great
    Observatories Origins Deep Survey (GOODS) fields. By employing
    publicly available X-ray and optical data and catalogues we discover
    68 type 2 AGN candidates.

File Summary:
--------------------------------------------------------------------------------
 FileName   Lrecl  Records   Explanations
--------------------------------------------------------------------------------
ReadMe         80        .   This file
table1.dat     90       47   Type 2 AGN candidates, HDF-N
table2.dat     90       21   Type 2 AGN candidates, CDF-S
table4.dat     90        3   Type 2 AGN candidates, UDF
--------------------------------------------------------------------------------

See also:
 J/AJ/126/539   : The Chandra Deep Fields North and South (Alexander+, 2003)
 J/ApJS/155/271 : Chandra Deep Field-South: Optical spectroscopy (Szokoly+ 2004)
         II/258 : Hubble Ultra Deep Field Catalog (UDF) (STScI, 2004)
         II/261 : GOODS initial results (Giavalisco+, 2004)

Byte-by-byte Description of file: table*.dat
--------------------------------------------------------------------------------
   Bytes Format Units   Label   Explanations
--------------------------------------------------------------------------------
   1- 19  A19   ---     GOODS   GOODS designation (JHHMMSS.ss+DDMMSS.s)
  22- 25  I4    ---     UDF     ? UDF designation (Cat. II/258, table 4 only)
  27- 29  I3    ---     A03     Alexander et al. (2003, Cat. <J/AJ/126/539>
                                sequential number, [ABB2003] CDFN NNN (table1)
                                 or [ABB2003] CDFS NNN (table2&3) in Simbad
  31- 33  I3    ---     S04     ? Szokoly et al. (2004, Cat. <J/ApJS/155/271>
                                sequential number, [SBH2004] XID NNNa in Simbad
                                (table2 only)
      34  A1    ---   m_S04     [a] Multiplicity index on S04
  35- 36  I2    h       RAh     Right ascension (J2000.0)
  38- 39  I2    min     RAm     Right ascension (J2000.0)
  41- 45  F5.2  s       RAs     Right ascension (J2000.0)
```

Attention!
Les tables du ReadMe peuvent etre différentes de celles accesibles via VizieR

Exemple de fichier de configuration .status

```
%! Catalogue Status
%----------------------
\cId{ J/A+A/530/A18/ }
%\cUsualName{   }       % Short Designation
%\cDic{tab}{colname}    % Link for VizieR-S
\cCenters{ f } %  ftp vizier Bejing India LaPlata Moscow NASA Tokyo zero_data
\cSimbad{0}             % Simbad Status 0=Not 1=Ids_only 2=Fully_accessible
%
\cAdded{ 28-Apr-2011 }  % When this file was added
\cBulletin{ --- }       % Bulletin number where catalogue announced
\cType{ MC }
\vizMerge{ table[3456] }{ recomb }{Case B effective recombination coefficients \
  for electron densities between 10^2^ and 10^5^cm^-3^ \
  {\em(tables 3--6 of paper)}}
\vizAddColumn{ table[3456] }{ logNe }{,2,3,4,5}{[cm-3]}{ -Tr}\
  {\ucd{PHYS_DENSITY_ELECT}Electron density (values 2, 3, 4, 5 in log)}
\vizExplain{ * }{}{+ \vizContent{timeSerie}(\wGraph{@{@catab}}\
   {}{Velocity plot})}
\vizSet{ * }{ logNe }{ fmt=1d dbtype=i1 type=I }
\vizDisplayColumns{ * }{ * }
vizUCD{ table[789] table1[01234] }{ Del del }{ =FIT_STDEV }
\vizUCD{ table[789] table1[01234] }{ Mult  }{ =AT_MULTIPLET_ID }
\vizSimbadName{ table5 } { table4.sim }
\vizPKlink{ table4 }{ Star }{ equivalent width }
\vizFKlink{ ew }{ Star }{ star general parameters }
```

Données en sortie (tables)

Données (tables) en sortie

- Les tables sont stockées en base de données.
- Les tables sont construites à partir du format ASCII CDS standardisé (excepté les grands catalogues).

Les grandes volumétries sont **aussi** stockées sous forme **binaire** :

- Ancien format : F.Ochsenbein
- Nouveau format : FX.Pinneau

Les métadonnées et données en sortie

- construites à partir des fichiers ReadMe et du fichier de configuration (.status)
- stockées en base de données dans un catalogue META (~30tables)
  - METAcat, METAtab, METAcol
  - METAfilter
  - METAcellxx
  - ....

Le métadonnées VizieR

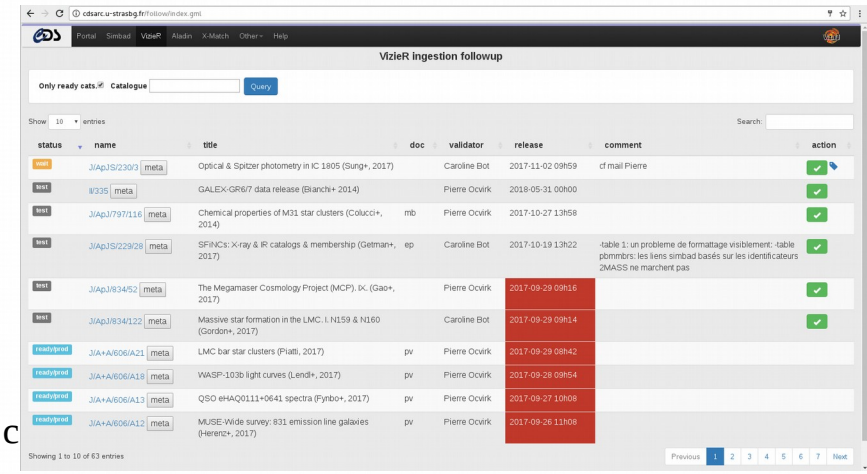
Table METAtab avec LaTeX dans la description des tables aujourd'hui séparée dans une nouvelle colonne “morexplain”

Idem pour la table METAcol

name	catid	tabid	explain	morexplain
I/332 table3	113320616	2	The log of IUE observations of V444 Cygni	IUE observations (\vMore{-source=VI/110/inescat \&-out.max=999\&-c=V444 Cyg,rs=5}{all}); illustrations of {\bf\aFile{J/AN/332/616/fig4.ps}{spectra at different phases (Fig.4)}}, and {\bf\aFile{J/AN/332/616/fig8.ps}{SiII line changes (Fig.8)}}
32 able2	113320616	1	Observational data for V444 Cyg	Observational data for V444 Cyg \vizContent{timeSerie} (\wGraph{J/AN/332/616/./table2}{P=4.212424}{light curves})
J/MNRAS /372/777/psr	73720777	1	Pulsar Survey: positions, flux densities, pulse widths, periods, dispersion measures and derived parameters	Pulsar Survey: positions, flux densities, pulse widths, periods, dispersion measures and derived parameters {\em(tables 1, 2 and 3 of paper)}
J/AJ/112 /407/stars	51120407	4	FBQS candidate list: stars	FBQS candidate list: stars {\em(table 4 of paper)}
J/AJ/112 /407/fbqs	51120407	1	FIRST Bright Quasar Catalog, part I	FIRST Bright Quasar Catalog, part I {\em(table1 of paper)}
J/AJ/112 /407/egal	51120407	2	FBQS candidate list: galaxies	FBQS candidate list: galaxies {\em(tables 2 and 3 of paper)}

L'ingestion des données, les outils

- La gestion du système de stockage: outil Unix: make_public, newcat ..
- L'extraction des données : getapj (pour les journaux de l'AAS)
- La standardisation des tables en format ASCII CDS et la génération du ReadMe
  - Outils partagés CDS, auteurs, AAS:
    - Anafile,acut
    - Pyreadme
  - Outils pour les auteurs: interface web http://cdsarc.u-strasbg.fr/vizier.submit/
- L'ingestion des tables dans la base de données: vizin (2V)
- La validation des catalogues (astronomes)



Vizier : un bilan

Bilan positif : VizieR créé en 1995

- Un architecture et un processus d'ingestion bien adapté pour la **préservation** : DSA
- Relation avec les producteurs de données:
  - les journaux : AAS (développement de getapj)
  - les agences: ESO, ESA, NASA (intégration des grandes volumétries)
    bonne gestion technique, format adapté et recherche active pour les volumétries de demain
- Ancré dans le VO , VizieR a adopté ET adopte les **standards du VO**
  → prohéminence de VizieR dans les registres VO

Mais aussi ...

- Technologie vieillissante qui n'est plus adaptée à son usage:
  awk, C, LaTeX (avec conséquence sur l'évolution)
- Fragilité du système d'ingestion assuré par une seule personne
- Un isollement VizieR comparé aux technologies en cours au CDS :
  echange de librairies, standards d'indexation
- Intégrer les nouveaux standards internationnaux

VizieR2 : but de la nouvelle version

Adopter les nouveaux standards

- **Domaine temporel** : ajouter le temps comme nouvel axe de recherche et proposer des sorties VizieR avec le futur modèle VO
- Passer en Full TAP (VO)
- Passer en UCD1+
- DOI/ORCID

Amélioration du système d'information

- Conservation de la précision originale en base avec formatage (?)
- Rationalisation de la gestion des données originales
- indexation globale positionelle avec prise en compte des époques
- Statistiques sur les data (demande pour TAP dans une future version?)
- (Revoir la gestion des mots clés ?)

Intégrer VizieR dans le framework du CDS

- Améliorer l'indexation spaciale : Qbox --> HEALPIX (Moc ?)
- interfacage grands catalogues / HiPS (?)
- partage des moyens (librairies) et ouverture du code

Evolutions technologiques

- Migrer la “recherche intuitive” vers une nouvelle technologie
- migration langage pour améliorer la maintenance

VizieR2 : déja en action ...

Le domaine temporel:

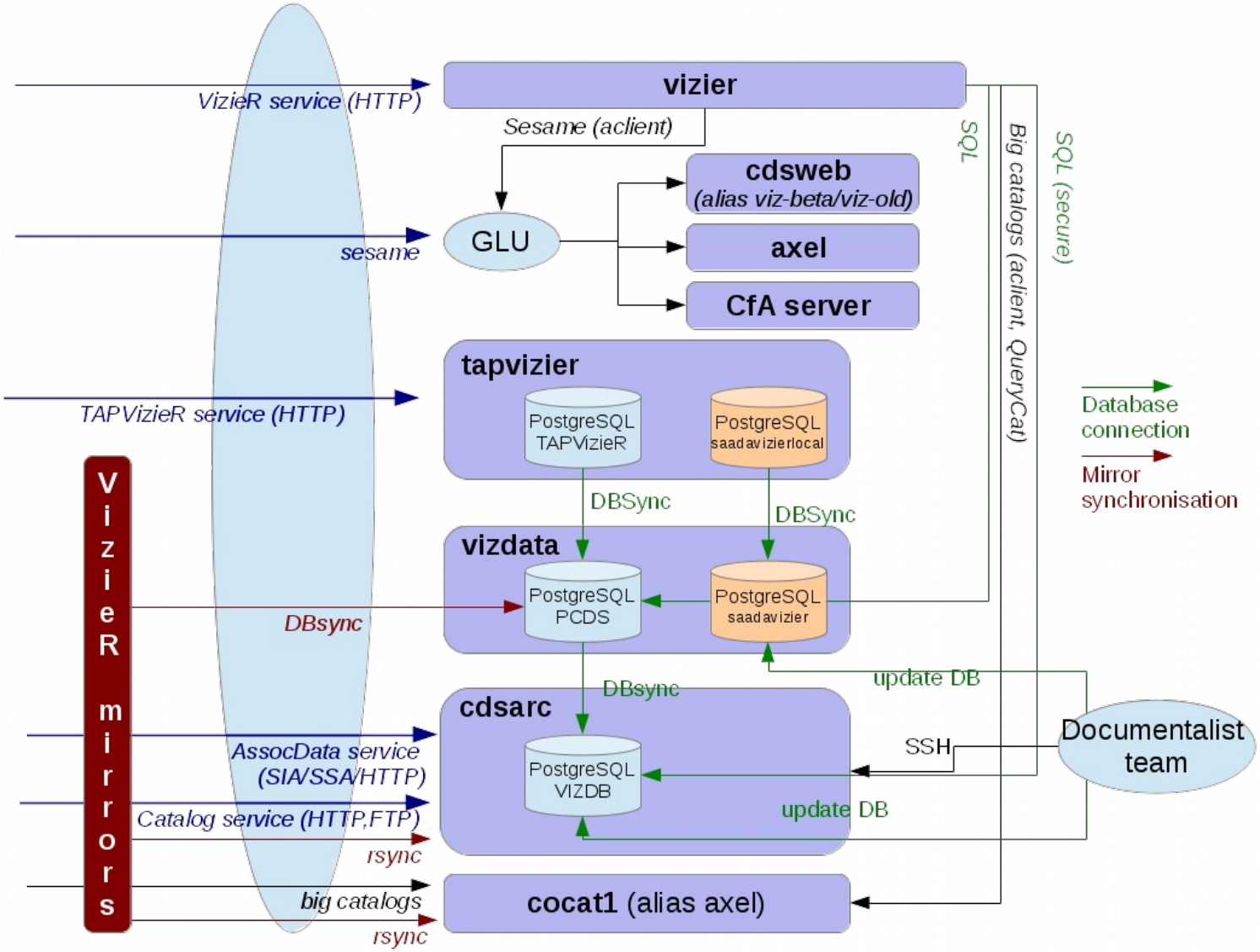
- Implication forte du CDS pour un standard VO
  → A.Neubot chaire domain interest group
  + plusieurs personnes impliquées au CDS notamment T.Boch pour le coté technique
- Réflexion sur le schéma de base de données pour integrer les métadonnées temporelles.
  Aujourd'hui dans le cadre de Gaia:
  -→ pas d'automatisation ni d'indexation et intervention manuelle en un 1er temps

Autres actions

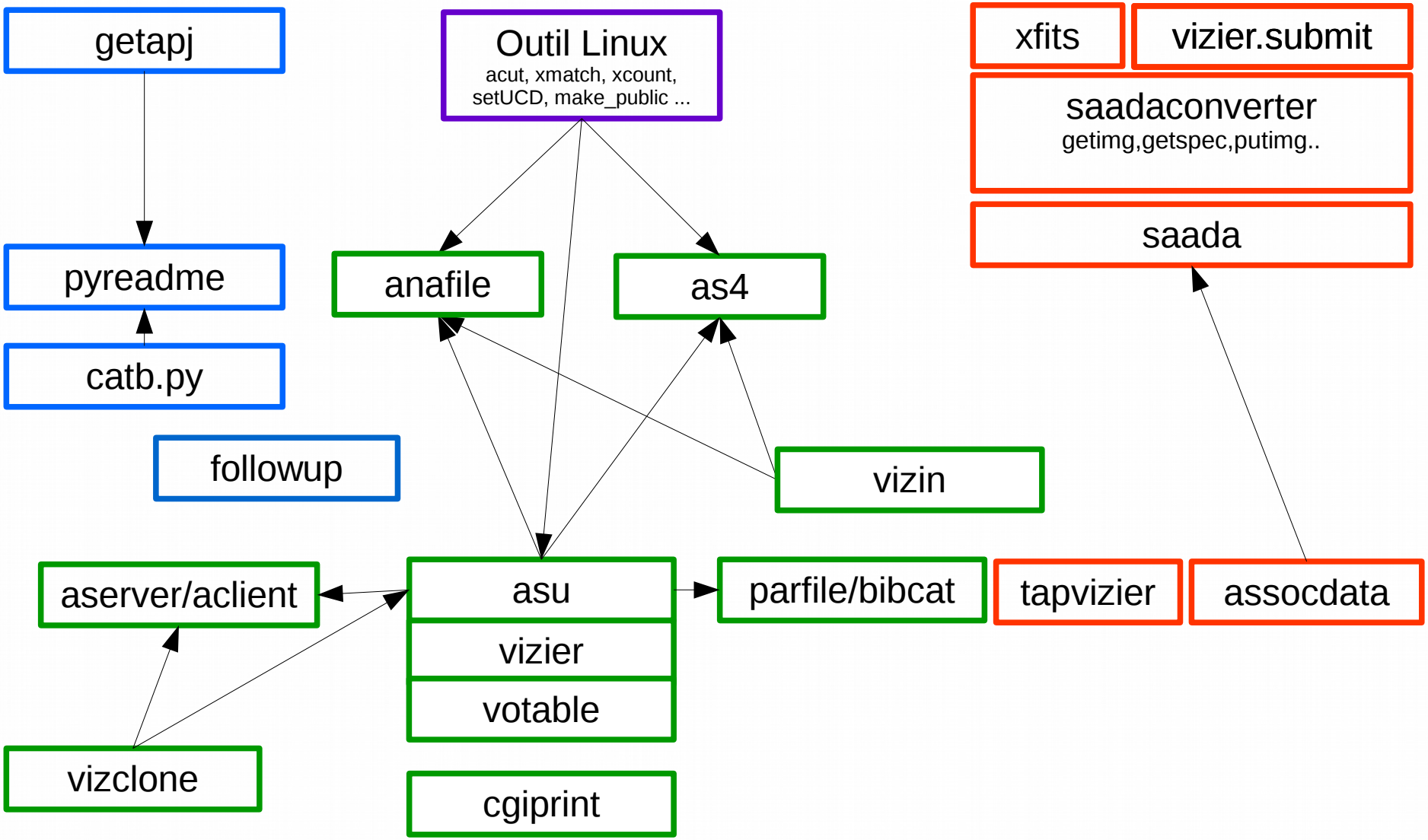
- Reflexion sur les DOI
- Recherche intuitive (stage avec L.Michel)
- Plus tard: recherche pour interroger des données binaire du CDS en SQL

Présentation technique

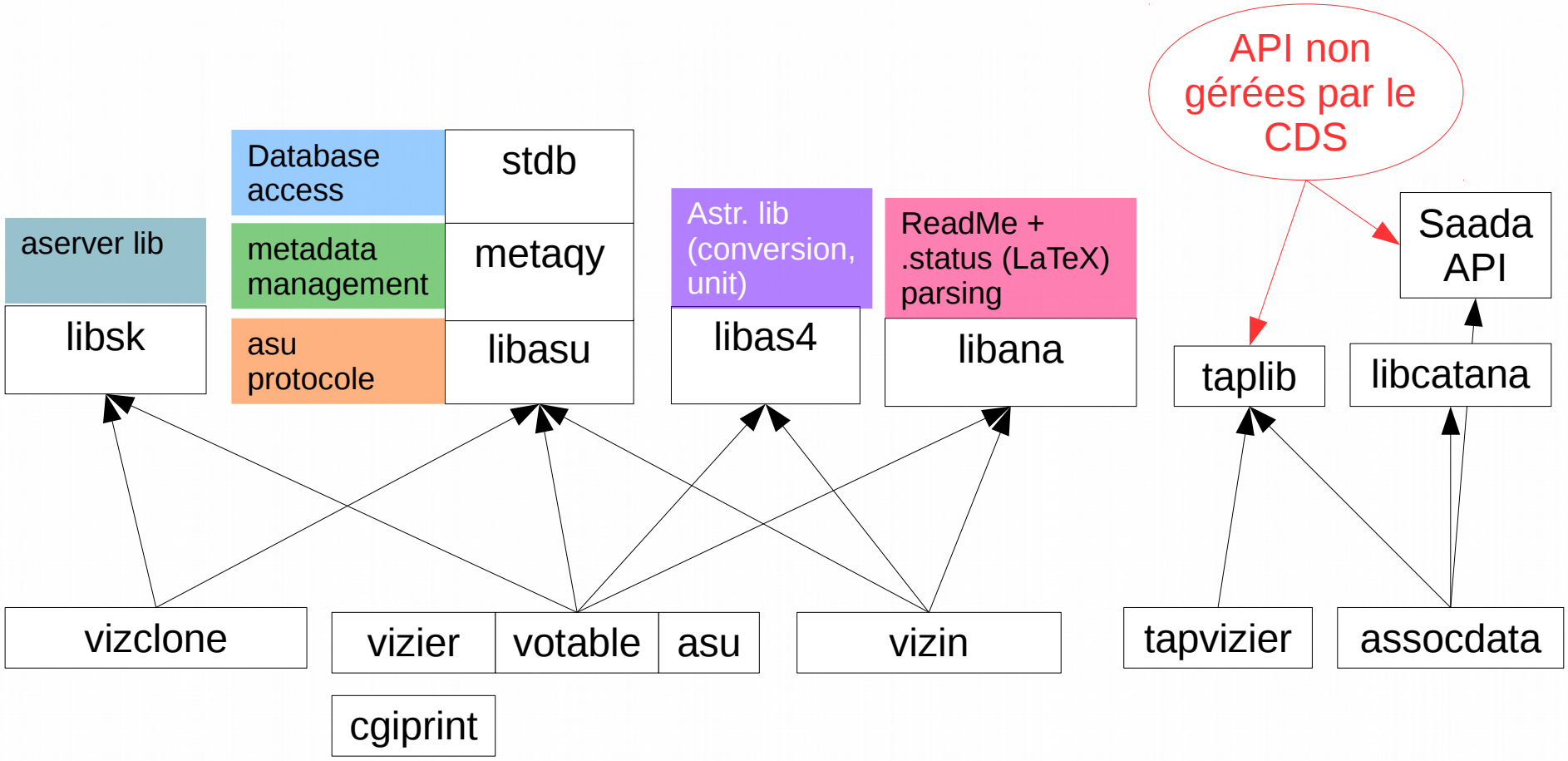
Architecture VizieR



Architecture VizieR



Architecture VizieR



Note: il existe aussi une indexation Kohonen – (sous forme de proc. stockée)

Indexation positionnelle VizieR

Indexation globale selon les Qbox (vizier classique)

- Modele hierarchique similaire à HEALPix mais avec une tesselation non uniforme (contrairement a HEALPix)
- Chaque emprunte de catalogue VizieR est ajoutée dans les tables METAcellxx selon un ordre 8 (~20minutes)

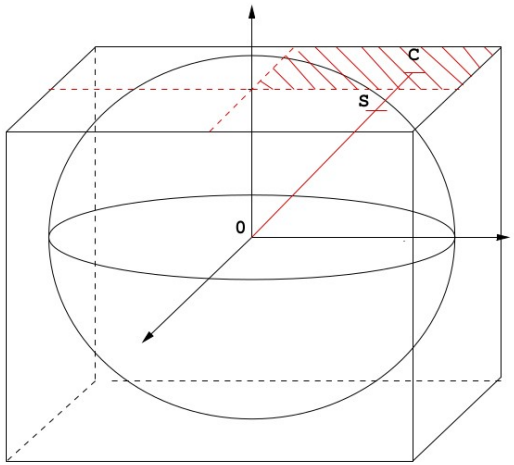
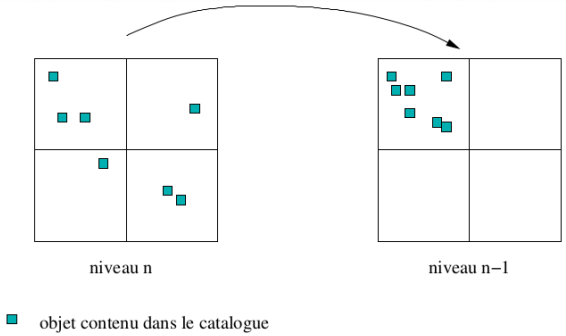
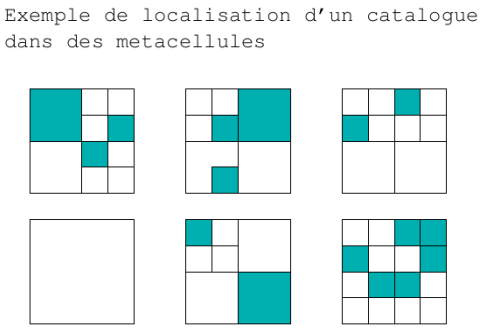


FIGURE 6 – Projection sur un cube





HEALPix (TAPVizieR)

→ Utilisation de la librairie H3C

Indexation positionnelle VizieR

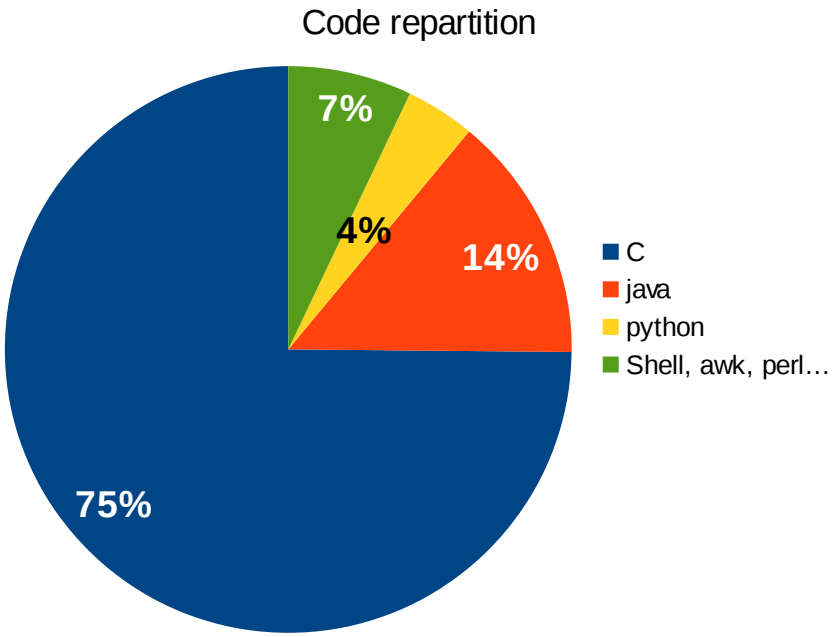
Etapes d'une interrogation VizieR selon une position:

1) Calcul des qbox recouvrant la zone recherche (box ou cone)
2) Recherche des tous les catalogues VizieR pour lequel le recouvrement Qbox contient des cellules qbox (1)
3) Pour chacun des catalogues, on affine la recherche au cone/box:
   → recherche si des données existent dans au moins l'une des tables:
   - 1er affinage: requete SQL qui borne les positions:
     type: select ... from ... where ra_min<ra<ra_max and de_min<de<de_max
   - 2d affinage: en memoire par le prog vizier

Architecture VizieR

Le code

language	Nb lignes
C	~265,000
Java	~50,000
Python	~14,000
Shell, awk,perl..	~25,000



lib	difficulté	TODO
stdb	+	x
meta	+++	X (partiellement)
asu	++	x
Vizier/votable	+	
as4	++++ ???	X (existe en java)
cgiprint	+++ ???	
parfile	???	x
vizin	++++	x
anafile	++	x

TODO ??.....

TODO ??.....

Todo ?? ...

Base de données:

- Rationalisation des schemas VizieR (viz1, viz2 ,....)
- Passer en Full TAP :
  - en utilisant le format grand catalogue (FDW)
  - (ou) tout en base ?
- UCD1+ : besoin de programmes similaire a ceux existant pour les UCD1
- Amélioration de la recherche positionnelle:
  - Migration Qbox → HEALPix : MOC service, METAcell, librairie M.Nullmeyer
  - Indexation des données : H3C/PgSphere ?

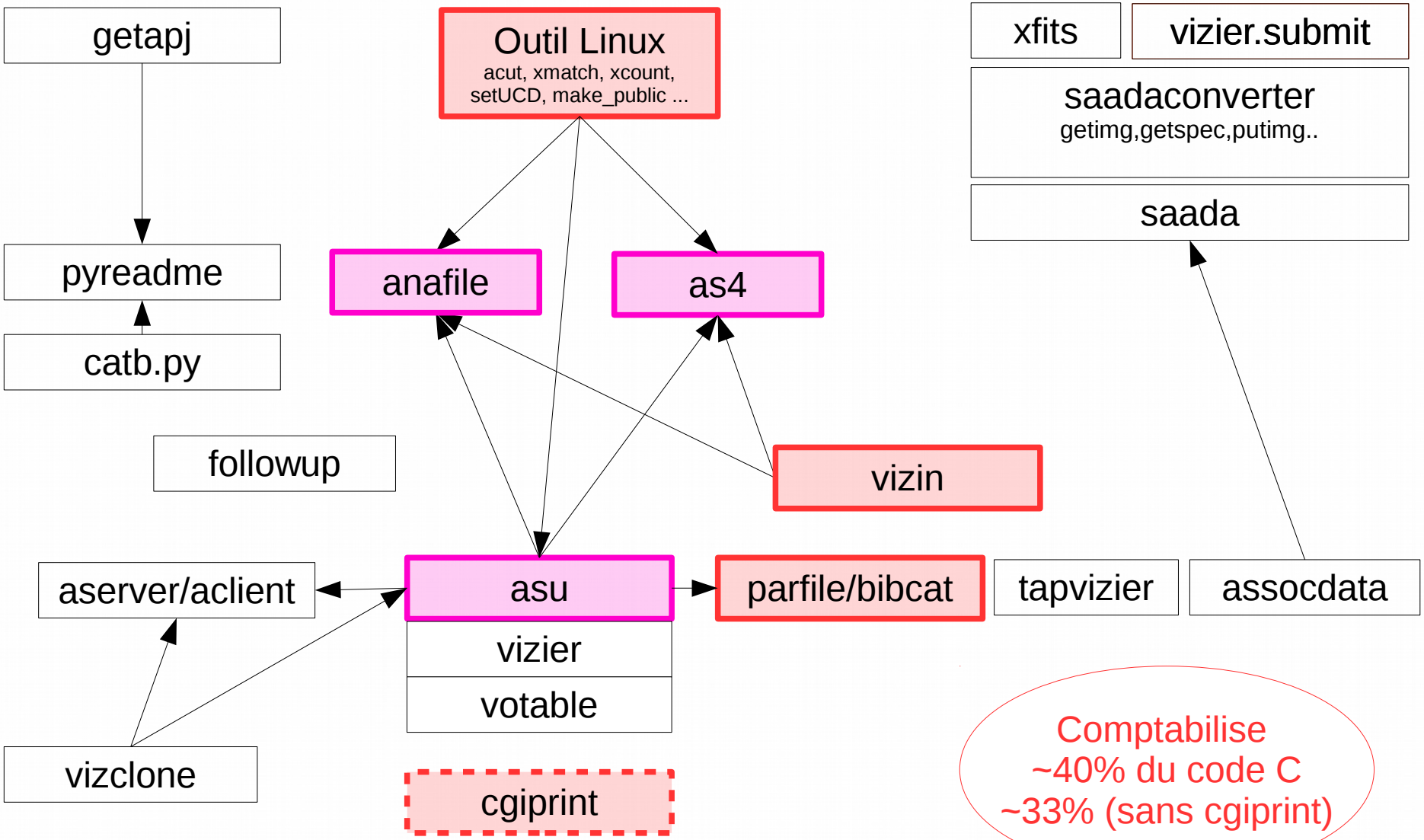
Todo ?? ...

Migration de programmes existants

- Le programme vizin d'ingestion
  - Lib parsing macro LaTeX du .status
    → JavaCC (Java) ou pyparsing (Python)
  - Parsing du ReadMe
  - Ordonanceur des commandes du .status et des métadonnées du ReadMe
  - Calculs de conversions des unités : temps, position, format, ... --> utilisation de Java as4 ?
  - Opération sur les tables : jointure, merge, links, ...
- Nouveau bibcat (basé sur parfile) pour l'indexation textuelle intuitive (Elasticsearch)
- SetUCD (?)

Éventuelement cgiprint ?

Code ciblé


getapj
Outil Linux
acut, xmatch, xcount,
setUCD, make_public ...
xfits
vizier.submit
saadaconverter
getimg,getspec,putimg..
saada
pyreadme
catb.py
anafile
as4
followup
vizin
aserver/aclient
asu
parfile/bibcat
tapvizier
assocdata
vizier
votable
vizclone
cgiprint
Comptabilise
~40% du code C
~33% (sans cgiprint)

Comparatif python - Java

Python

- En pleine extension, notamment en astronomie
- De nombreuses librairies : numpy, astropy !
  => dépendance externe ! ??
- Adapté au travail sur les tables
- Possibilité d'interfacage avec le C : libana
- Outil de développement : pycharm, debugger
- J'aime Python !
- Possibiliteéde passage en “douceur” en interfacant des code C (j'ai jamais fait)

Java

- Robuste
- Existe depuis 1995
- Très utilisé au CDS
- Librairies catana, as4, HEALPix, UWS ... ? (xmatch, ...) ??
  éventuelement tapLib si migration du service VizieR
- Outil de développement : debugger , Eclipse (par contre lourd)
- Rupture brutale avec l'existant _ pas de partage de lib avec le service vizier