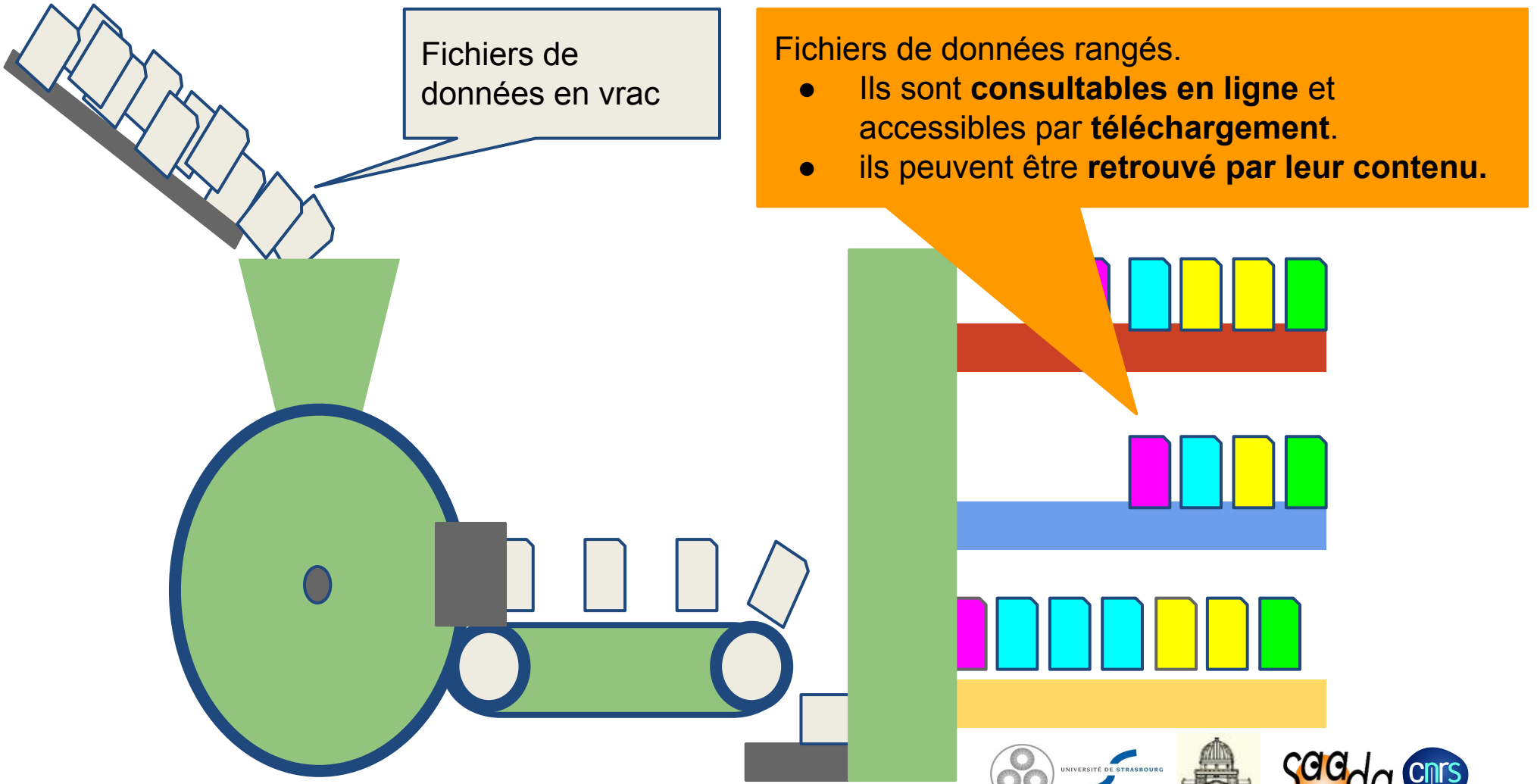
http://saada.unistra.fr

laurent.michel@astro.unistra.fr

Une machine à ranger et a indexer des données astro



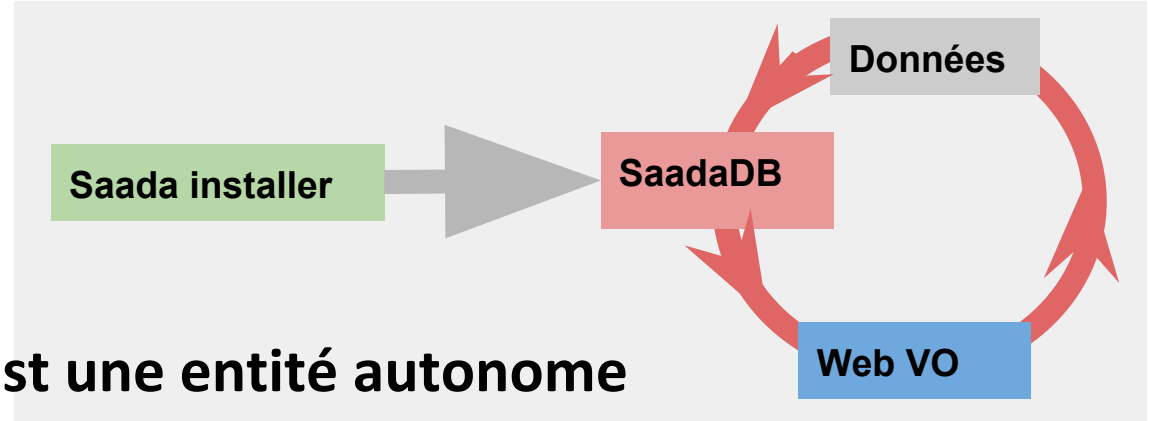
Saada en quelques mots

- ✓ **Construction automatique d'une base de données à partir de fichiers de données (FITS, VOTable …).**
  - ➢ **Pas de code** à écrire
  - ➢ Archivage de **données hétérogènes**
  - ➢ Hébergement de **collections multiples**
- ✓ **Extraction et archivage des données**
  - ➢ Extraction des mots clés lus dans les fichiers chargés
  - ➢ Rangement des données de manière à faciliter la réponse aux requêtes des utilisateurs
  - ➢ Le fichier d'origine reste accessible
- ✓ **Outils de gestion et de publication des données**
  - ➢ Possibilité d'établir des **liens persistants** entre les données.
  - ➢ Marquage à la main des méta-données (UCD, unités…)
  - ➢ Accès par le biais de l'interface Web ou de protocoles VO

Saada en quelques mots

✓ Le principe

- ➢ Un installeur met en place un canevas de base de données, la SaadaDB) dans lequel l'utilisateur organiser le stockage des données.



✓ Une fois créée, la SaadaDB est une entité autonome

- ➢ Plus de lien avec l'installeur
- ➢ Pas de connexion avec un serveur central

✓ Constituants de la SaadaDB

- ➢ Le système de stockage (PosgtresQL, MySQL ou SQLite)
- ➢ Le data loader (Interface graphique scriptable)
- ➢ L'interface Web (RIA basée sur JQuery)
- ➢ Le repository (répertoire de travail et d'archivage de fichiers)

Saada en quelques mots

- **Saada**
  l'installeur s'appelle Saada comme le projet (choix malheureux).

- **SaadaDB**
  La base de données construite à partir du canevas a son propre nom. Elle est toutefois référencée dans les documents comme une SaadaDB.

- **Produit**
  Ensemble de mots clés/valeurs
  - Ensemble vide: fichier plat (flatfile)
  - Ensemble non vide: fichier structuré (FITS ou VOTable)

- **Format d'un produit**
  Ensemble des mots clés qui lui sont associés (typage inclus)

- **Collection hétérogène**
  Ensemble de produits dont au moins un a un format différent d'un autre

Saada en quelques mots

✓ Une SaadaDB peut être vue comme:

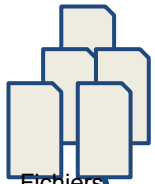
- ➢ Une **base relationnelle** dont les tables contiennent des données extraites de fichiers par le data-loader. Cette base peut être utilisée par n'importe quelle application ayant un connectivité avec le monde des SGBDRs.
- ➢ Une (multi) **ressource OV** permettant d'accéder à des données via SIA, SSA, CS ou TAP.
- ➢ Une **application Web**

✓ Les données d'une SaadaDB peuvent être sélectionnées par

- ➢ Des requêtes simples: position, mots clés
- ➢ Des requête ayant une saveur *fouille de données*

```
Select all XMM detections
located near ABEL 426
     and having a flux greater than 1e-13
     and matching at least one Simbad source
          having an obj_type containing the string `Radio'
```

Cycle de vie des données



Fichiers chargés

Data Loader

Classes Java générées

API Java bas niveau

Base SQL

SaadaDB

Accès Java aux entités Saada
collections, classes, instances d'objets
Possibilté d'accès depuis une autre application Java

Accès Java aux entités SQL
Tables colonnes

Accès libre aux entités SQL
Tables colonnes
Possibilté d'accès depuis n'importe quelle autre application

Les concepts clés

✓ Les collections

- ➢ Les données sont réparties dans des collections définies par l'utilisateur
- ➢ Une collection est un conteneur abstrait doté d'un nom
- ➢ L'affectation d'une collection à un ensemble de fichiers à charger est entièrement libre

✓ 5/6 catégories de données

- ➢ Les données sont organisées en 5 catégories qui correspondent au grandes classes de données astronomiques

catégorie	*Type de fichiers*	*Type d'extension*
FLATFILE	Any	no
MISC	FITS, VOTable	Any
SPECTRUM	FITS, VOTable	Table or image
IMAGE	FITS	Image
TABLE + ENTRY	FITS, VOTable	Table

2 niveaux d'archivage

✓ Stockage des données natives

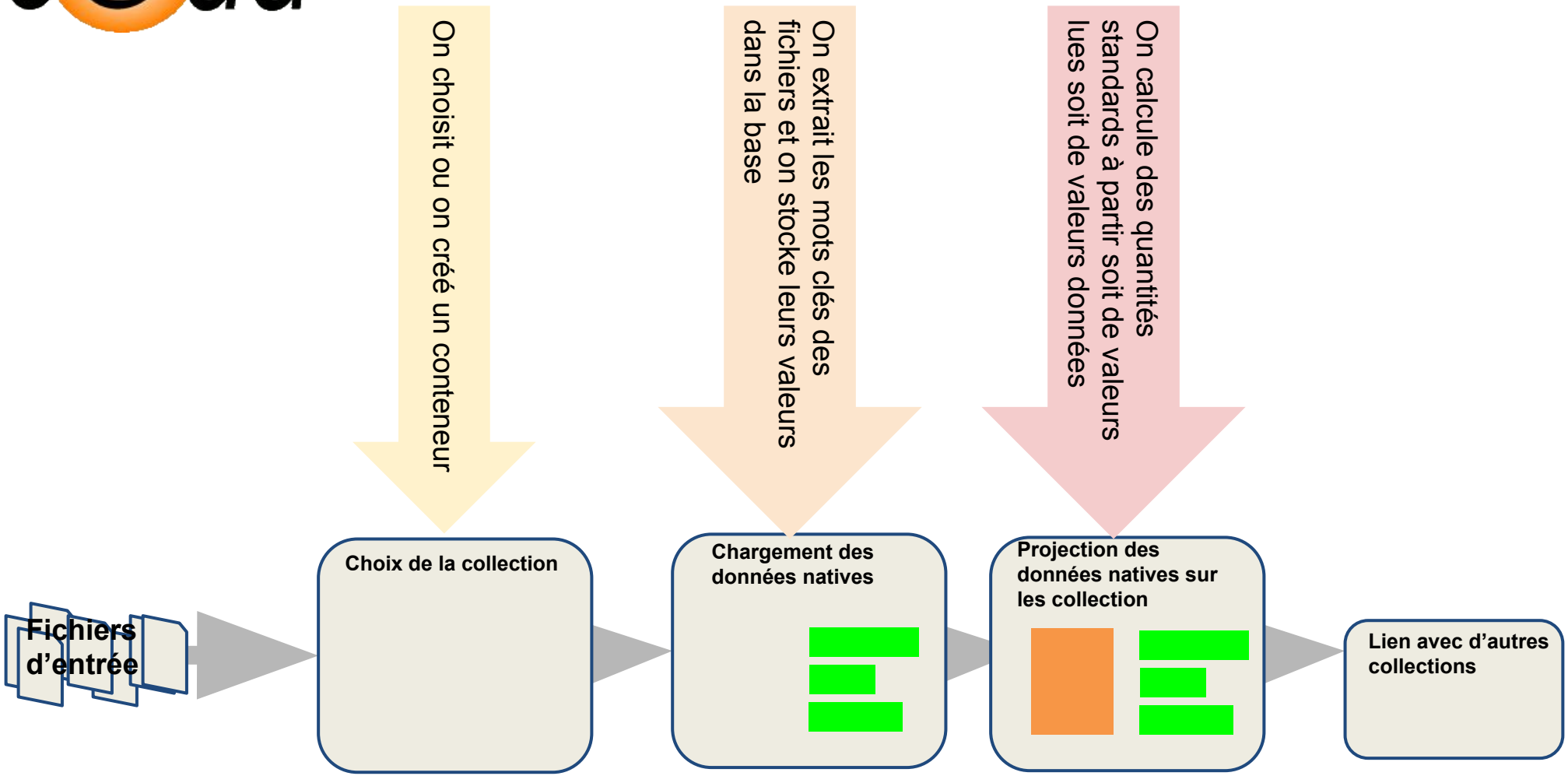
- Les valeurs extraites des fichiers sont **stockées telles quelles**
- Elle restent accessibles en lecture ou utilisables comme critères de sélection

✓ Création d'une interface commune

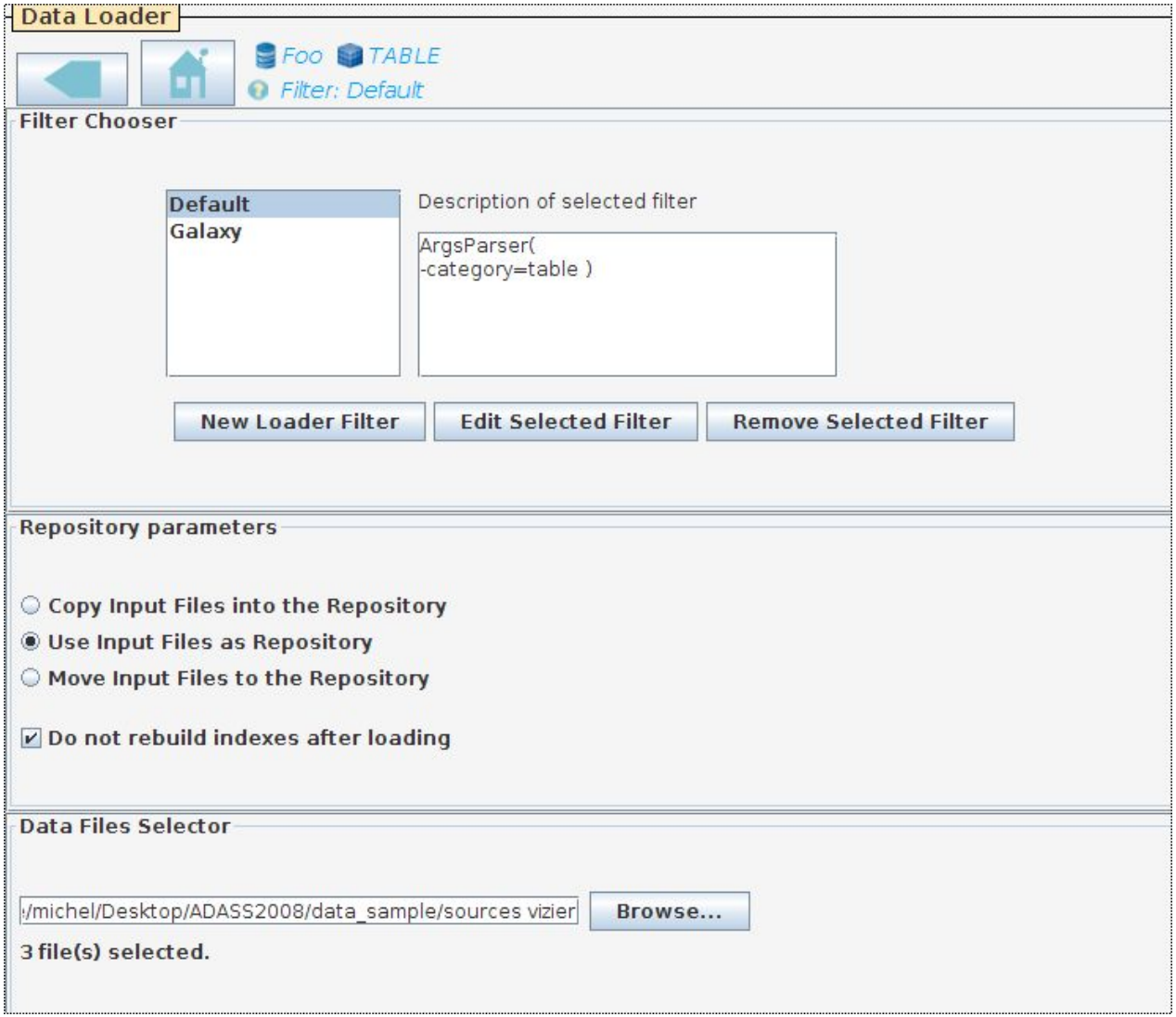
Rupture entre 1.8 et 2

- Saada calcule un certain nombre de valeurs caractéristiques qui sont également exposées
- Le format de ces valeurs est **commun à toute la base** (nom, type unité…).
- Elles couvrent les axes de caractérisation (espace, temps, énergie, observation)
- Elles permettent de faire des **requêtes** sur des collections de données **initialement hétérogènes**
- L'utilisateur peut ajouter ses propres colonnes à cette interface commune
- Ces colonnes sont calculées automatiquement, mais l'utilisateur peut fournir des règles
- Cette interface commune est nommée **données de niveau collection**

Chargement des données



Démo 1



Chargement images par défaut

Classes de données natives

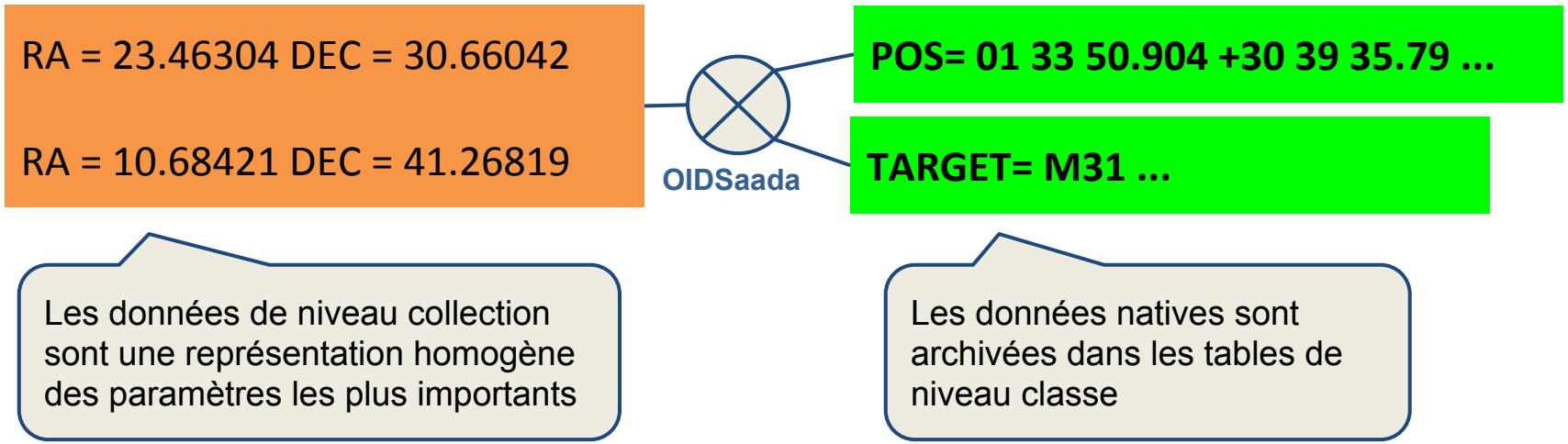
- ✓ **Les données natives sont regroupées dans des classes**
  - ➢ Une classe contient un ensemble d'enregistrements de **même format**
  - ➢ Elle est implémentée à la fois sous la forme d'une **classe Java** et d'une **table SQL**
  - ➢ Elle est attachée à une catégorie dans une collection
  - ➢ Les classes sont utilisées pour délimiter le périmètre des requêtes

- ✓ **Deux modes de classification**
  - ➢ Le mode **classifier** (mode par défaut): tous les fichiers ayant le même format (même jeu de mots clés) sont rangés dans une même classe.
  - ➢ Le mode **fusion**: tous les fichiers sont rangés dans la même classe quelque soit leurs formats.
    - ✓ Les colonnes non affectées restent nulles
    - ✓ Les conflits de types sont résolus par transtypage (boolean-> string)
  - ➢ Le choix du mode de classification relève de l'utilisateur

Les collections de données

✓ Structure interne d'une collection

- ➢ Une collection est découpée en 6 conteneurs: un par catégorie + ENTRY
- ➢ Pour chaque catégorie:
  - ✓ Une **table** par pour les données de **niveau collection**
  - ✓ Une **table** par classe de **données natives**
- ➢ Les tables de niveau **classes** et **collections** sont **jointes** par un identifiant universel de Saada (OIDSaada)



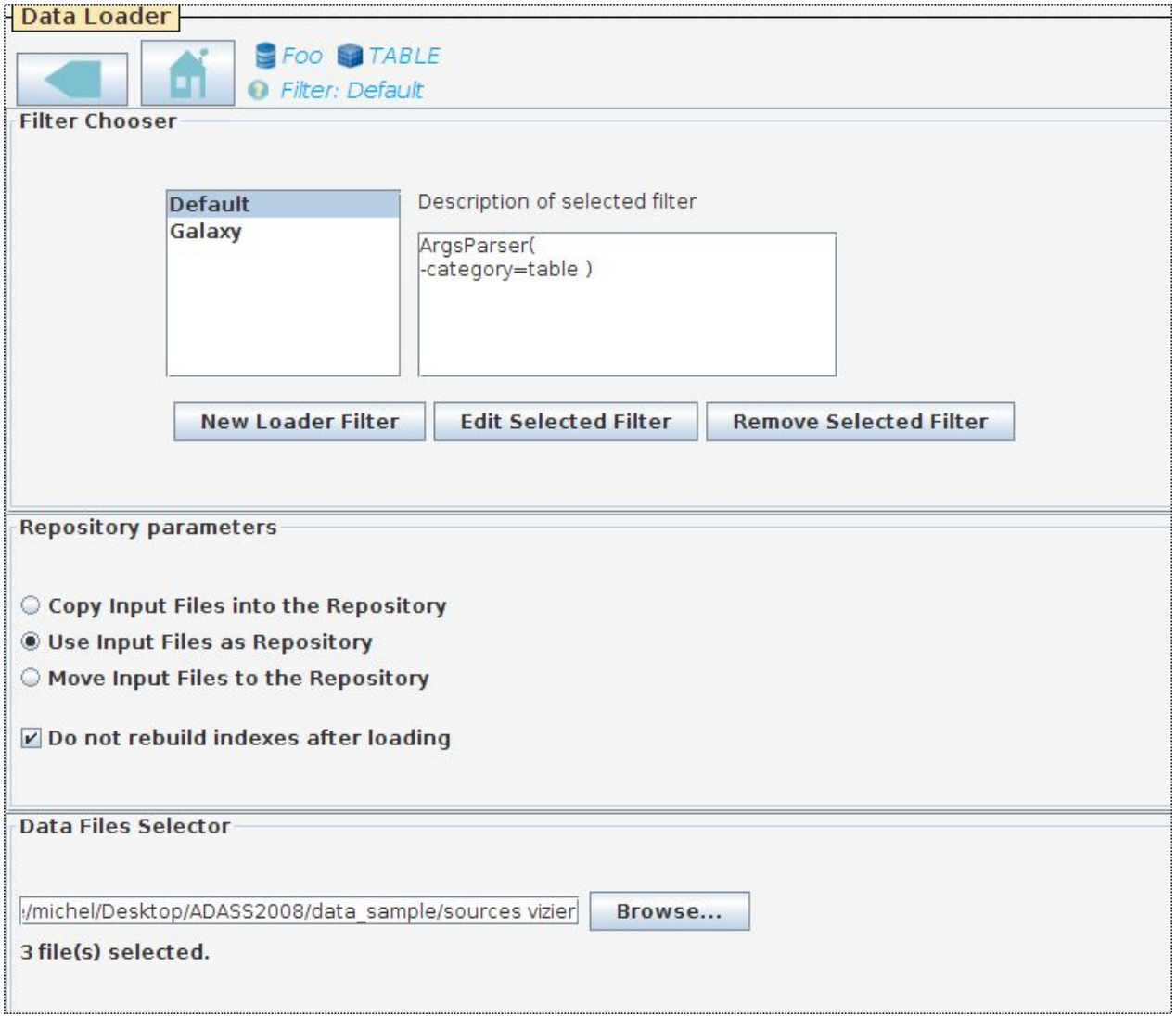
Calcul des donnés collection

- ✓ **2 modes d'affectation:**
  - ➢ **Détection automatique** par le data loader: recherche basée sur les noms de colonnes et les UCDs
  - ➢ **Règles utilisateur:** pour chaque colonne l'utilisateur peut fournir soit le mot-clé natif correspondant soit une valeur constante.
- ✓ **3 modes de priorité:**
  - ➢ L'utilisation des deux modes est régulé par un ordre de priorité défini par l'utilisateur
    - ✓ ONLY: on applique uniquement la règle utilisateur
    - ✓ FIRST: (defaut) on applique d'abord la règle utilisateur, puis la recherche automatique en cas d'échec.
    - ✓ LAST: on applique d'abord la recherche automatique puis la règle utilisateur en cas d'échec.

-position='M33' : La valeur de la position (niveau collection) pour tous les produits appliquant cette règle sera la position de M33 .

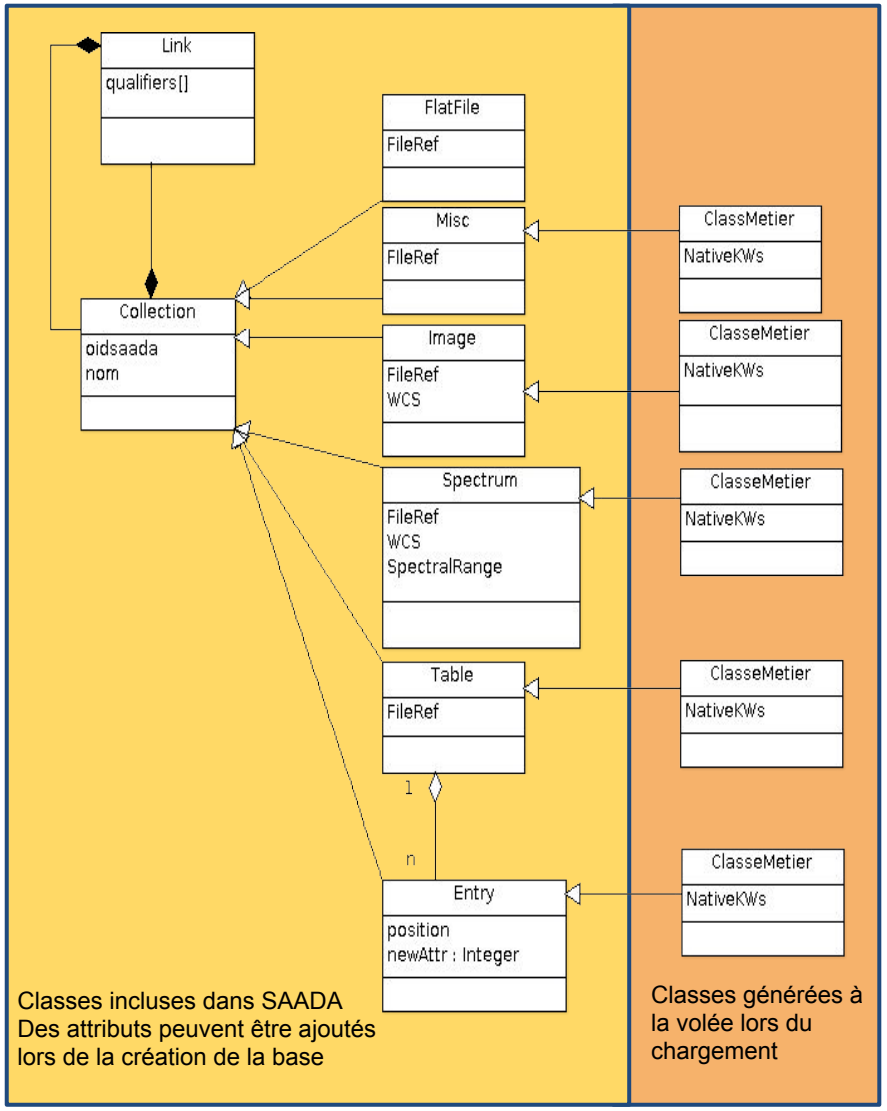
-position=RA,DEC: Tous les produits pour lesquels cette règles est appliquée auront pour position (niveau collection) les valeurs des mots clés RA et DEC

Démo 2

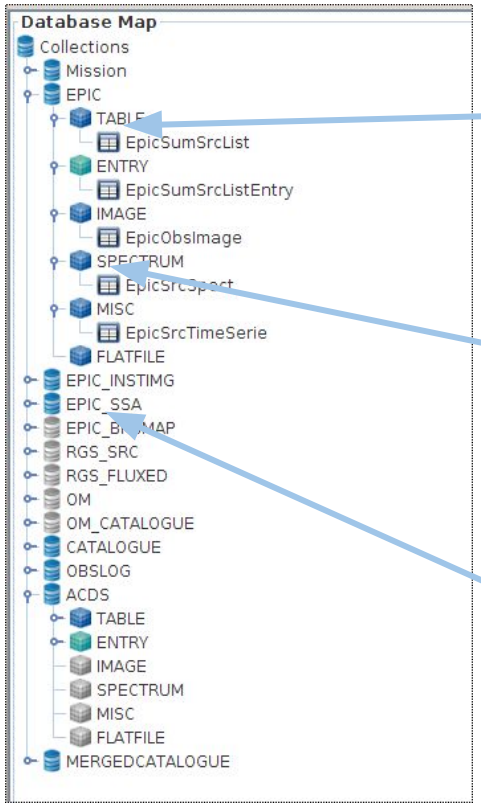


Chargement images une classe une position

Le modèle de données
UML vs IKEA



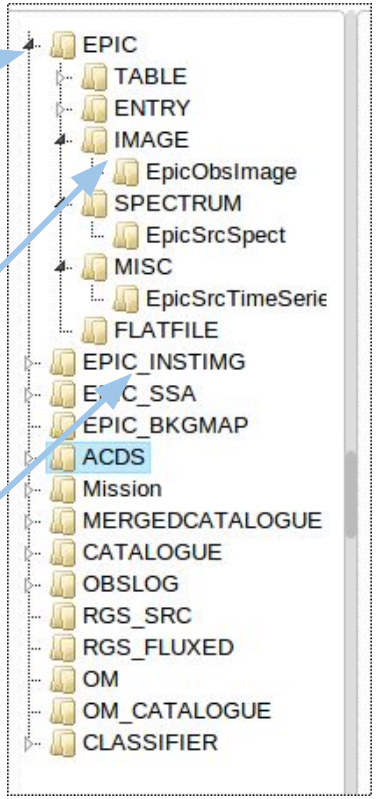
Vues sur le modèle de données



collection
Conteneur de données

catégorie
Données calculées par Saada suivant un modèle commun à l'ensemble de la base

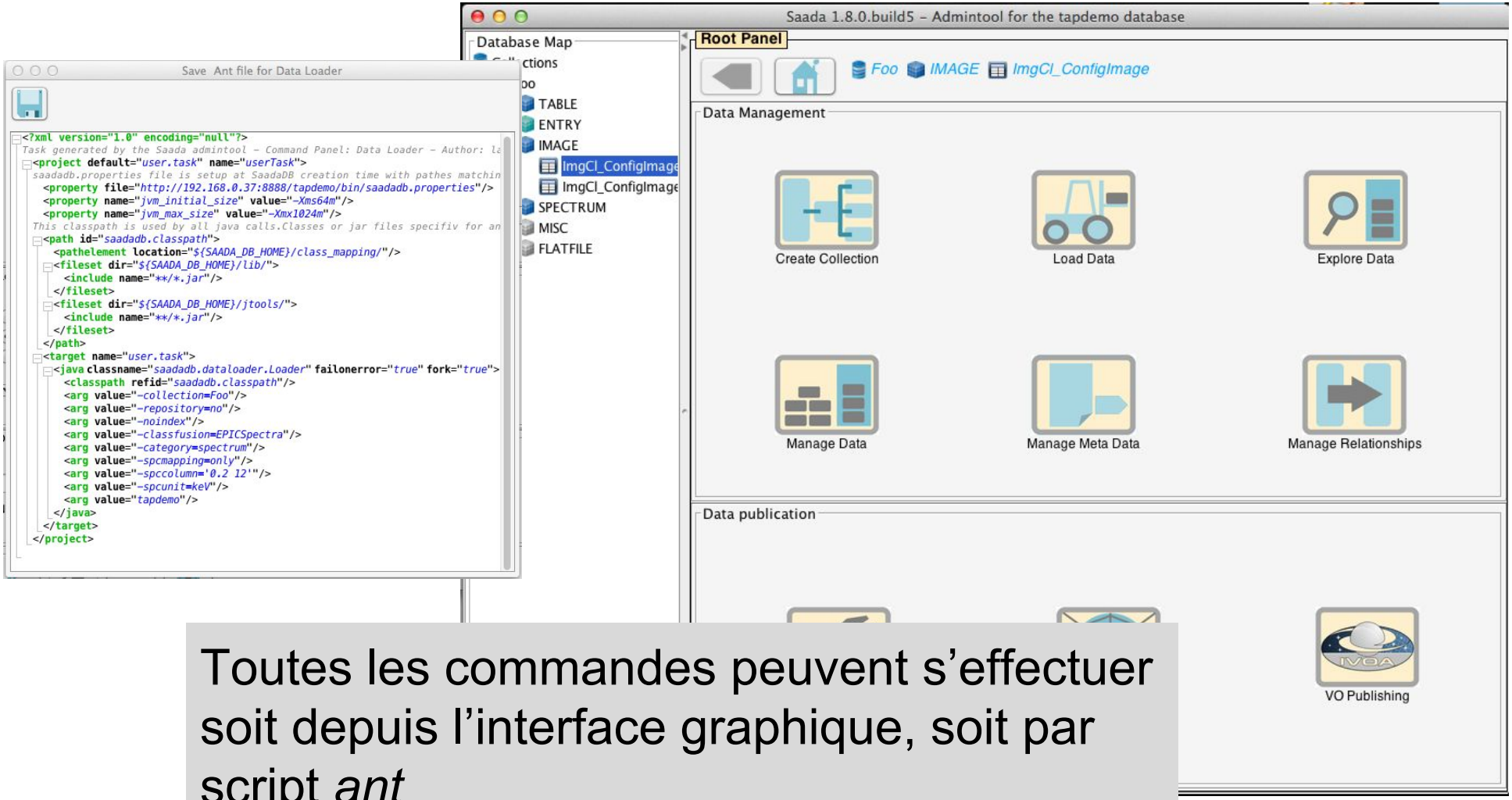
classe
Données natives des produits chargés



Depuis l'outil d'administration

Depuis la page Web

Gérer une SaadaDB



Toutes les commandes peuvent s'effectuer soit depuis l'interface graphique, soit par script *ant*

L'interface WEB

✓ Une application JEE

- Servlet + JQuery
- Déploiement sur Tomcat depuis l'outil d'administration
- Doit être redéployée après chaque modification des classes de données

✓ Principales fonctions

- Navigation dans le contenu de la base
- Éditeur de requêtes SaadQL
- Éditeur de requêtes S*P
- Connexion SAMP
- Téléchargement par lots via un Caddie

L'interface WEB

Localisation des données affichées

Gestion du Caddie

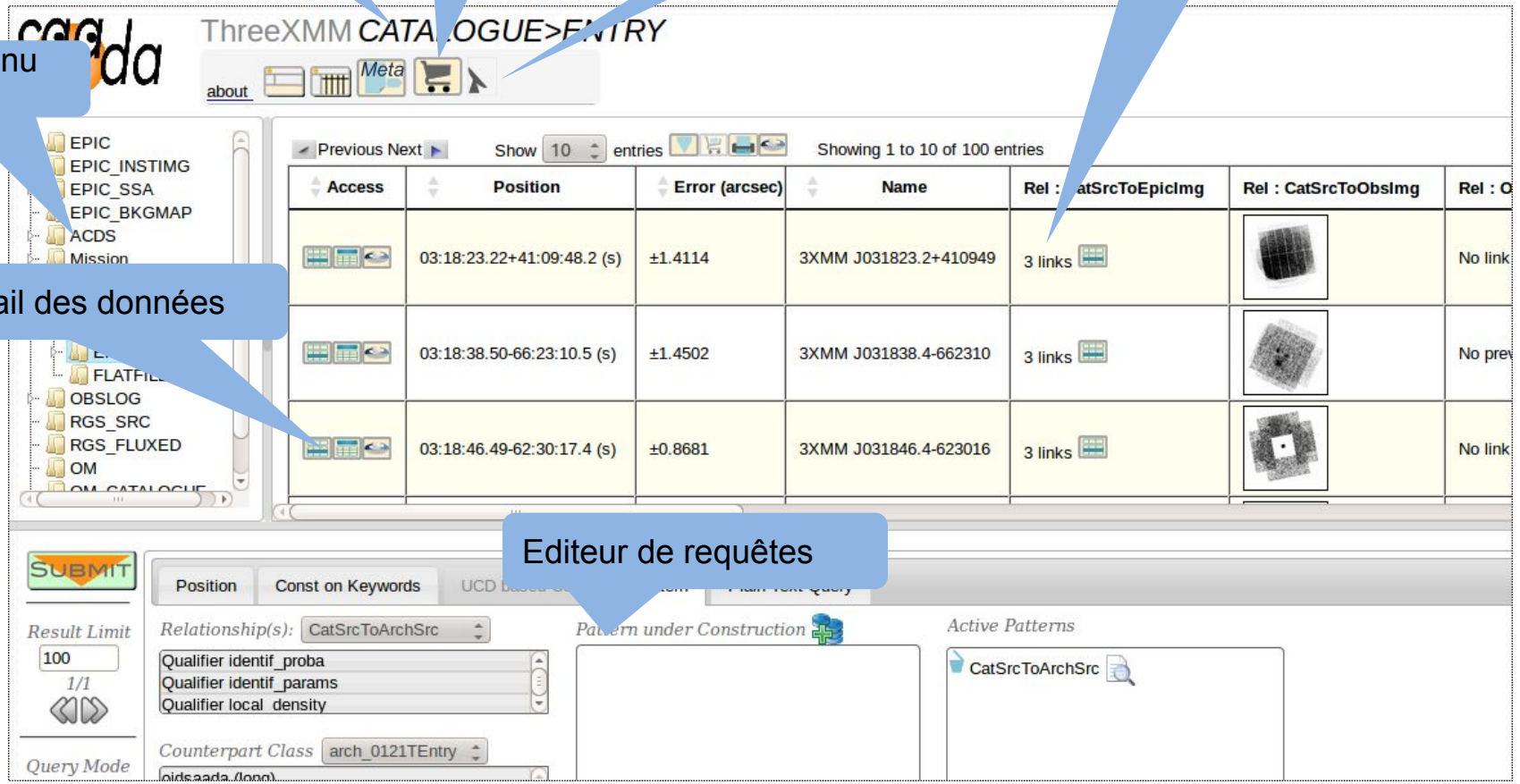
Connecteur SAMP

Sélection de données

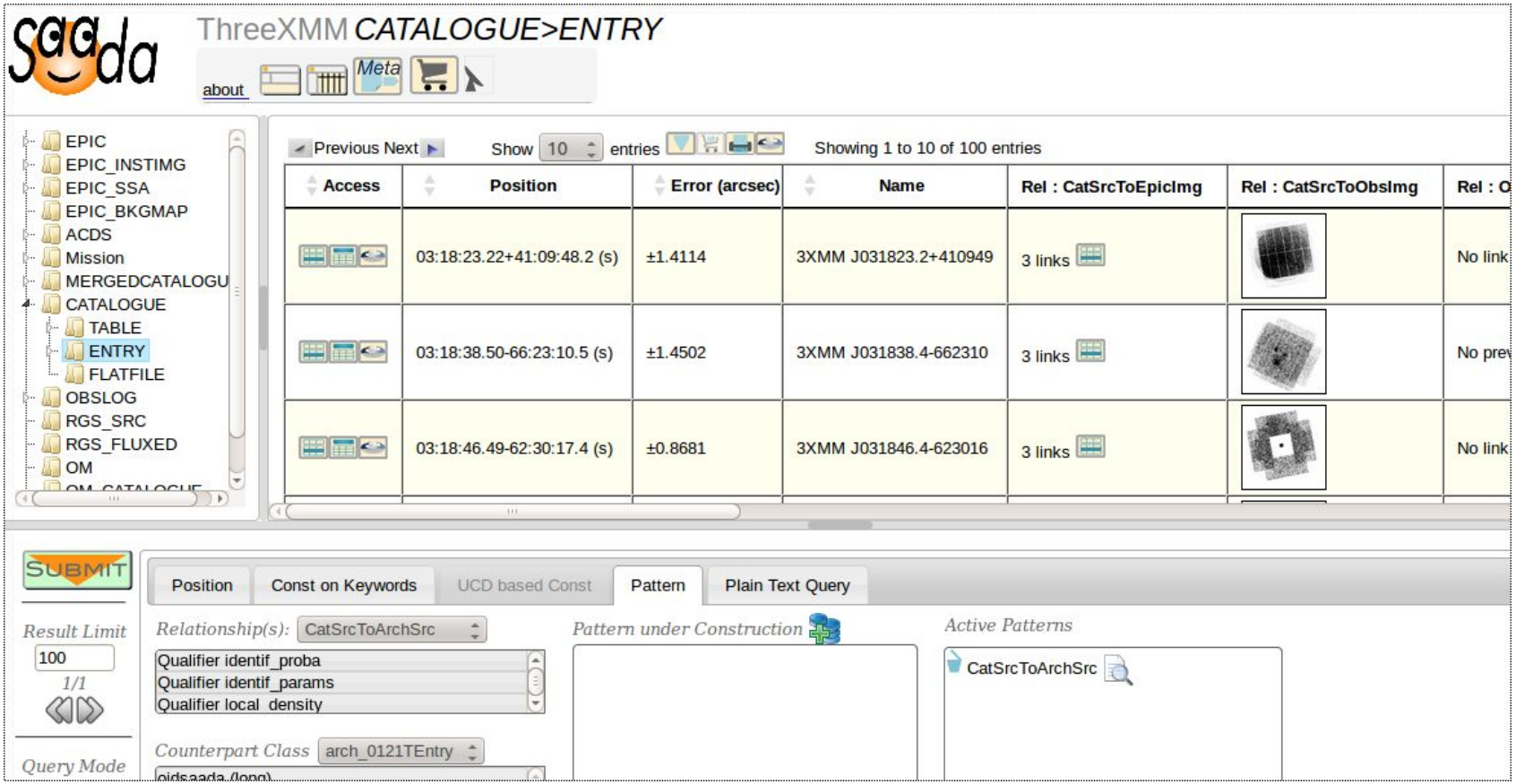
Arbre du contenu de la base

Accès aux détail des données

Editeur de requêtes



Démo 3



Interface Web

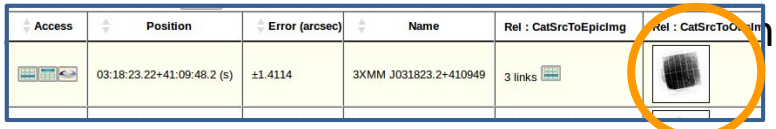
Les relations Saada

✓ Des liens permanents qualifiés

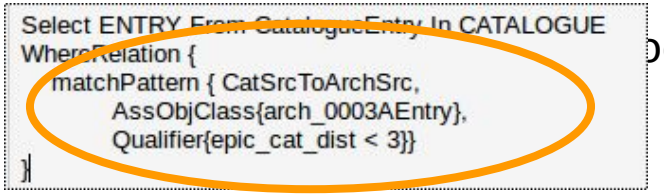
- Saada put gérer des liens persistants reliant des données de deux catégorie/collection
- Ces liens peuvent être qualifiés par des valeurs numériques (distance p.e.)
- Les relations ont un nom utilisable dans des requêtes

✓ Utilisation des relations

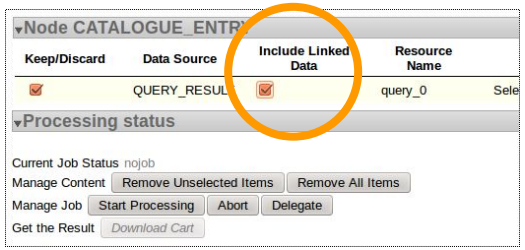
- Intégration des données associée sur la page Web

- 

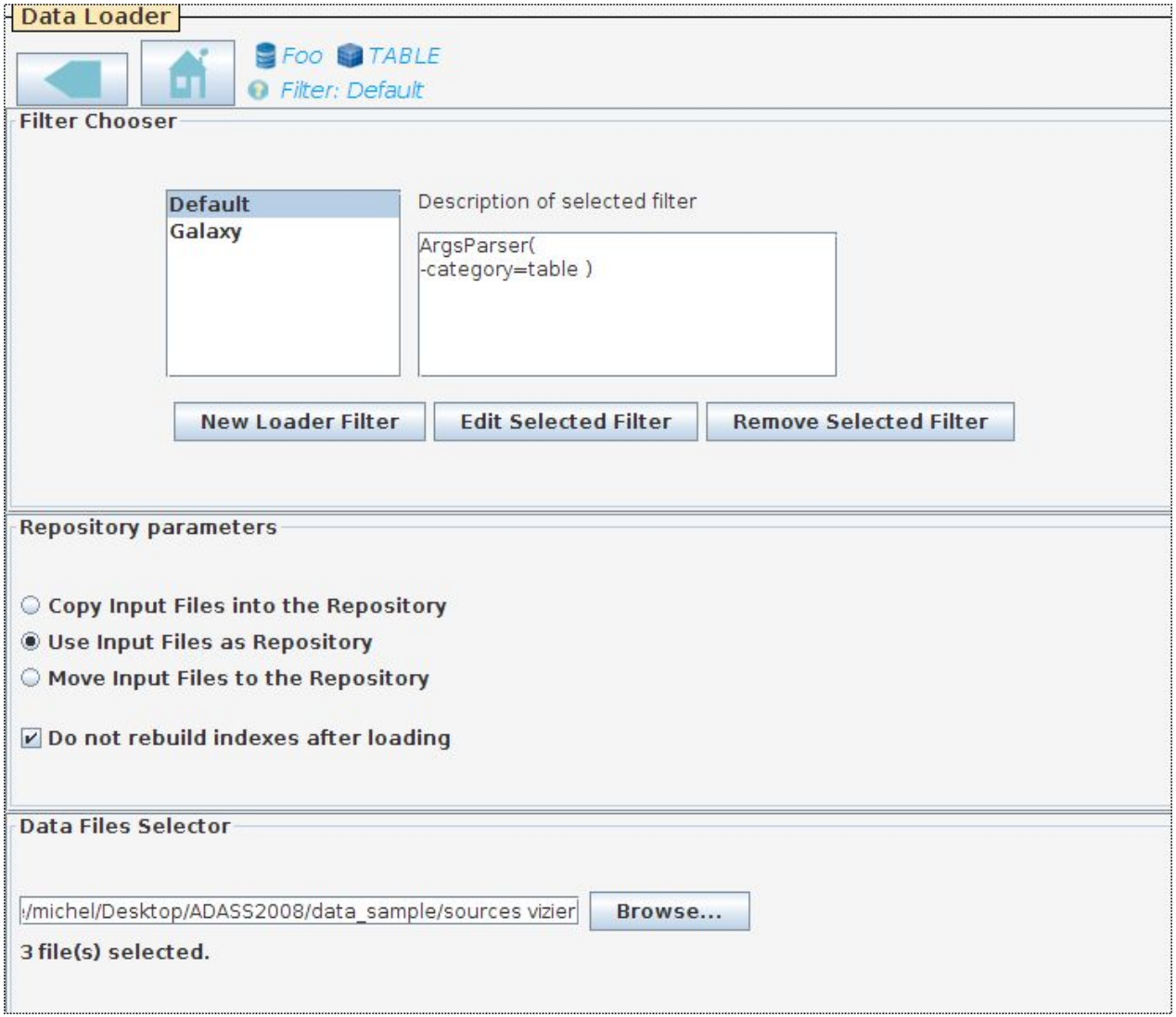
nées associées

- 

ociées dans le Caddie



Démo 4



Relation readme

SaadaQL

✓ Saada a son propre langage de requêtes: SaadaQL

- ➢ Les requêtes endossent les périmètres tracés par le modèle de donnée de Saada collection/catégorie/classe
- ➢ Les requêtes ne peuvent retourner que des OIDSaada. C'est à l'API de rechercher le contenu des objets

Select *CATEGORY* **From** *CLASSES* **In** *COLLECTION*

✓ 4 types de contraintes

`WherePosition`	Liste des positions de recherche
`WhereAttributeSaada`	Filtre SQL s'appliquant au données de niveau collection et aux données de niveau classe si la porté de la requête se limite à une seule classe
`WhereRelation`	Liste de filtres appliqués aux vecteurs formés par les liens partant de la collection sur laquelle porte le requête
`WhereUCD`	Filtre exprimé en utilisant les UCDs

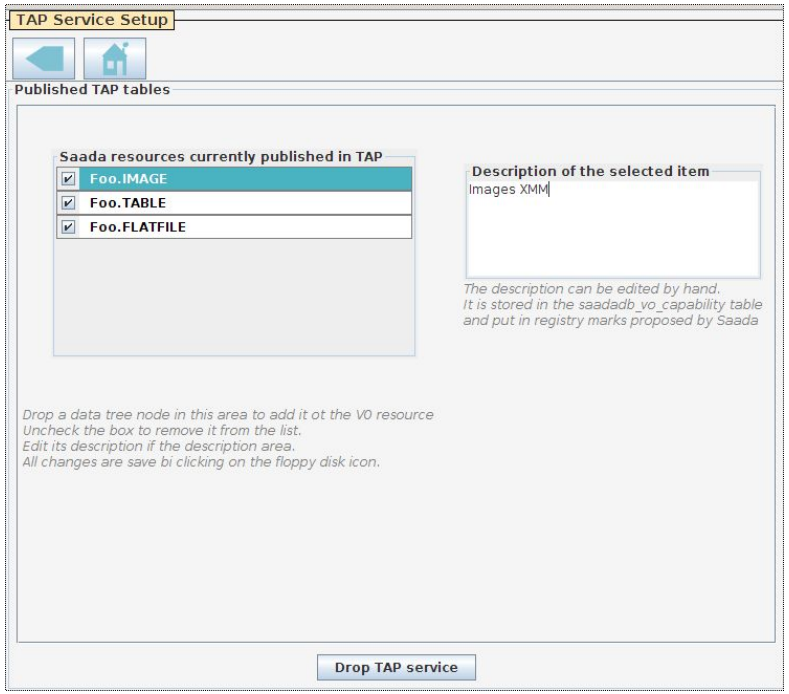
UNe requête complexe

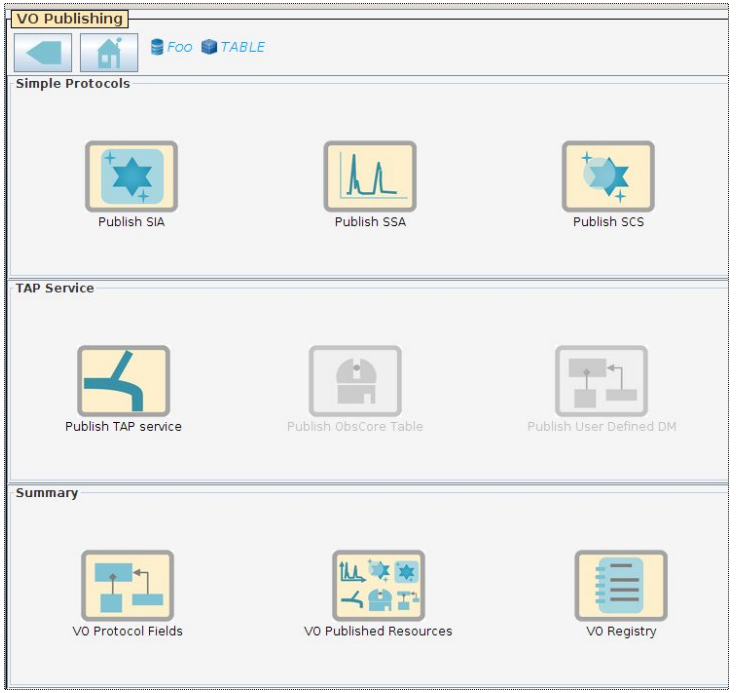
```
Select ENTRY  From CatalogueEntry  In CATALOGUE
WherePosition {
      isInCircle("M33",1, J2000, FK5)
     }
WhereAttributeSaada {
     XCAT_OBSERVER  = 'Isaac Newton'
     and namesaada = 'NGC598'  and
     _obs_id > 123000
     }
WhereUCD {
     [phot.flux] < 1e-14 [erg/s/cm2]
     }
WhereRelation {
     matchPattern { CatSrcToArchSr
      ,  AssObjClass{ Arch_8034AEntry }
          ,  AssObjAttSaada{_bj < 20 }
          ,  Qualifier{ epic_cat_dist < 1)
         }
       matchPattern {CatSrcToArchSrc
         , Cardinality > 1
         , AssUCD{ [src.class] = 'QSO  [none]}
         , AssObjAttSaada{ namesaada like '%q%"}
        }
}
```

Services VO

✓ Publication des ressources

- ➢ Les ressources S*P n'ont pas besoin d'être publiées pour être accédées via une requête OV. La fonction de publication de la SaadaDB permet juste de créer une description pour le registry.
- ➢ Les tables visibles par TAP doivent par contre être explicitement déclarées

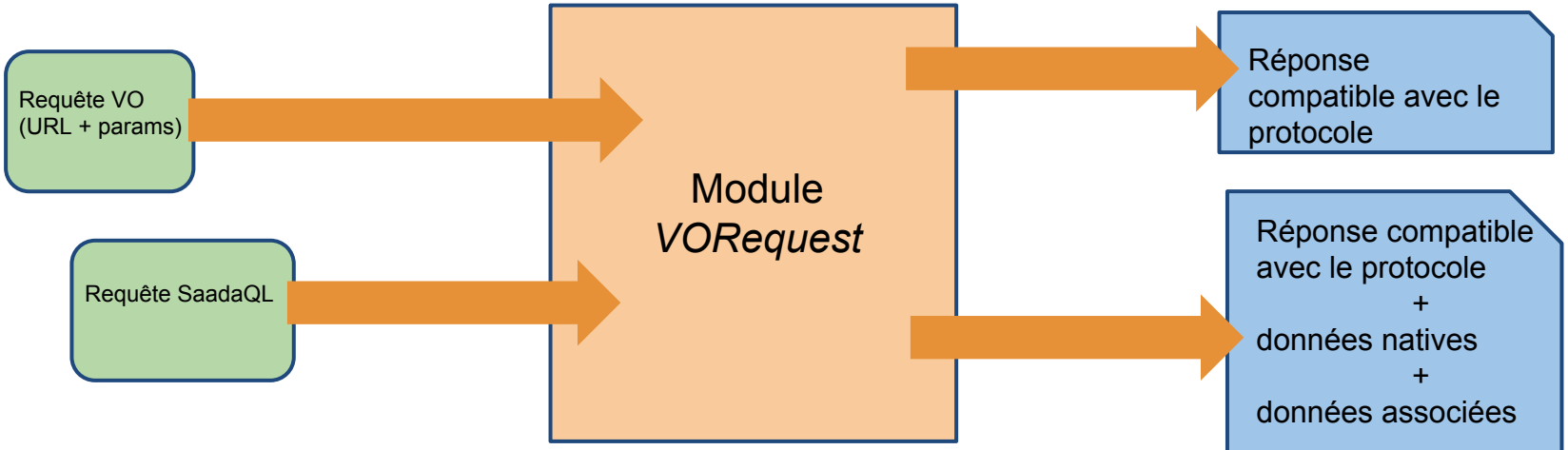


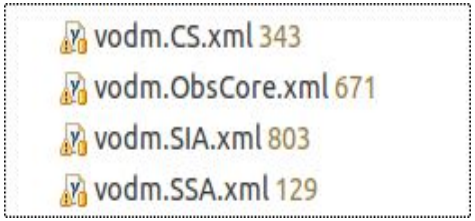


Services VO

✓ Flexibilité des réponses

➢ Le contenu des réponses S*P est contenu dans une VOTable. Il peut être adapté en éditant ce fichier.



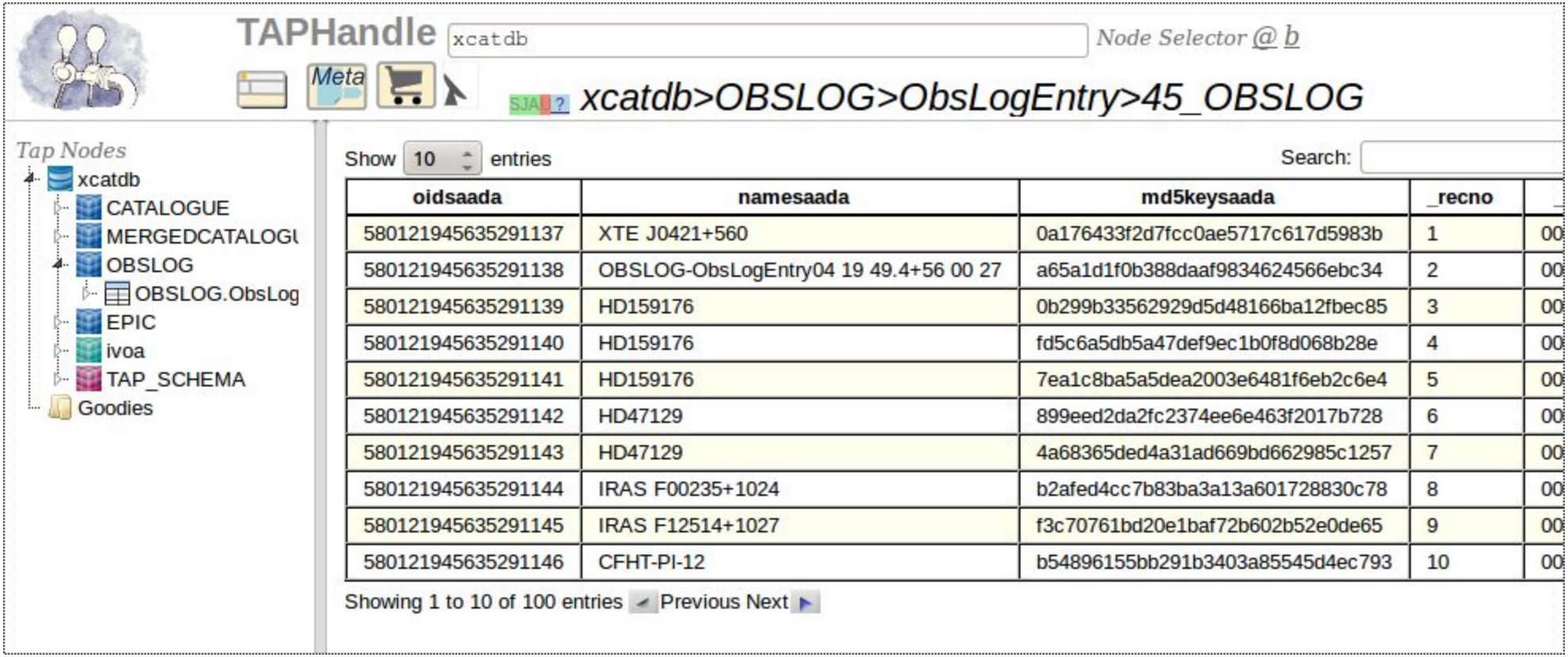


Modèles inclus dans la distribution

Description
du
modèle
(VOTable)

```
<?xml version="1.0" encoding="UTF-8"?>
<VOTABLE xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:noNamespaceSche
<RESOURCE type="VO Protocol Result Description">
 <DESCRIPTION>Saada VO service mapping</DESCRIPTION>
 <INFO name="Cone Search default" value="OK"/>
 <TABLE name="Results">
  <PARAM name="dm" value="Cone Search default" datatype="char" arraysize="*" />
  <FIELD name="group"          datatype="char" arraysize="*" />
  <FIELD name="name"           datatype="char" arraysize="*" />
  <FIELD name="ucd"            datatype="char" arraysize="*" />
  <FIELD name="utype"          datatype="char" arraysize="*" />
  <FIELD name="type"           datatype="char" arraysize="*" />
  <FIELD name="arraysize"      datatype="int" />
  <FIELD name="unit"           datatype="char"  arraysize="*"/>
  <FIELD name="hidden"         datatype="boolean" />
  <FIELD name="default_value"  datatype="char" arraysize="*" />
  <FIELD name="desciption"     datatype="char" arraysize="*" />
  <FIELD name="requ_level"     datatype="char" arraysize="*">
    <DESCRIPTION>MAN=1 REC=2 OPT=3</DESCRIPTION>
  </FIELD>
  <DATA>
  <TABLEDATA>
    <TR>
     <TD>DataID</TD>
     <TD>unique_id</TD>
```

Démo 4


TAPHandle
xcatdb
Node Selector @ b
Meta
SJAU ?
xcatdb>OBSLOG>ObsLogEntry>45_OBSLOG
Tap Nodes
xcatdb
CATALOGUE
MERGEDCATALOGU
OBSLOG
OBSLOG.ObsLog
EPIC
ivoa
TAP_SCHEMA
Goodies
Show 10 entries
Search:
oidsaada
namesaada
md5keysaada
_recno
58012194563529113 7 XTE J0421+560 0a176433f2d7fcc0ae5717c617d5983b 1
58012194563529113 8 OBSLOG-ObsLogEntry04 19 49.4+56 00 27 a65a1d1f0b388daaf9834624566ebc34 2
58012194563529113 9 HD159176 0b299b33562929d5d48166ba12fbec85 3
58012194563529114 0 HD159176 fd5c6a5db5a47def9ec1b0f8d068b28e 4
58012194563529114 1 HD159176 7ea1c8ba5a5dea2003e6481f6eb2c6e4 5
58012194563529114 2 HD47129 899eed2da2fc2374ee6e463f2017b728 6
58012194563529114 3 HD47129 4a68365ded4a31ad669bd662985c1257 7
58012194563529114 4 IRAS F00235+1024 b2afed4cc7b83ba3a13a601728830c78 8
58012194563529114 5 IRAS F12514+1027 f3c70761bd20e1baf72b602b52e0de65 9
58012194563529114 6 CFHT-PI-12 b54896155bb291b3403a85545d4ec793 10
Showing 1 to 10 of 100 entries
Previous
Next

Évolution de Saada

✓ Saada au CDS

- ➢ Fin 2013, le CDS propose d'utiliser Saada pour gérer les données attachées à Vizier
  - ✓ Stockage
  - ✓ Interface Web
  - ✓ Services OV
- ➢ Version 2.0Beta attendue pour la fin de l'année (intégration en cours)

✓ Adaptations nécessaires

- ➢ Utilisation de ObsCore pour les données de niveau collection.
- ➢ Outil interactif d'aide au mapping des données de niveau collection
- ➢ Utilisation d'expression pour le mapping:

em_max	NOT SET	NOT_SET	
em_min	NOT SET	NOT_SET	
em_res_power	NOT SET	NOT_SET	
facility_name	XMM	BY_KEYWORD	kw TELESCOP detected by name
healpix_csa	9223372036854775807	BY_SAADA	
instrument_name	EPIC	BY_KEYWORD	kw INSTRUME detected by name
naxis1	648	BY_SAADA	
naxis2	648	BY_SAADA	
o_calib_status	3	BY_VALUE	Infered from detected units Infered from detected units

```
em_max = RESTFREQ + CRVAL1 + (NAXIS1 - CRPIX1)*CDELT1
```

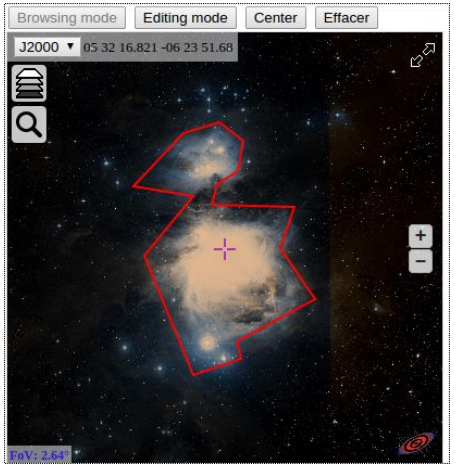
Évolutions de Saada

✓ Chargement des données

- ➢ Règles de mapping incluant des expression arithmétiques
- ➢ Pré-visualisation du mapping des données de niveau collection
- ➢ Amélioration des performances

✓ Requêtes

- ➢ Support des requêtes par régions
- ➢ Éditeur graphique de régions



✓ Interfacec OV (en cours)

- ➢ Mise à niveau des protocoles simples
- ➢ TAP Upload
- ➢ SIAP V2
- ➢ DataLink (transposition OV des relations Saada dans les VOTables réponses)