Retour d'expérience sur les méthodes développées par le centre de données astronomiques de Strasbourg



Gilles Landais (CDS)



Gilles Landais - Retour d'expérience sur les méthodes développées par le centre de données astronomiques de Strasbourg

Contenu des données labellisées

- VizieR
- Aladin archive des images en base de données, interrogeables par l'application Aladin ou les services de l'Observatoire Virtuel.

Simbad ne fait pas partie des données labellisées par le DSA

- → La base de données Simbad ne peut être vue comme une archive
 - Simbad est une base de données vivante construite au jour le jour à partir des informations contenues dans les articles publiées
 - L'identification des objets dans Simbad n'est pas figée (une étoile observée peut s'avérer avec l'amélioration des télescopes être une étoile double).

Présentation du CDS

Le CDS (centre de données de Strasbourg) existe depuis 1972, ses objectifs initiaux sont :

- Collecter des données utiles sur les objets astronomiques
- Enrichir les données en les évaluant de façon critique et en les combinant
- Distribuer les résultats à la communauté internationale
- Conduire des recherches utilisant les données

Pour ce faire:

- Le CDS dispose d'un équipe de documentalistes qualifiés et aidés par des astronomes (le CDS ~30 personnes)
- Le CDS développe des outils et mène des actions de R&D
- Le CDS participe à des projets dans son domaine de connaissances:
 CoRoT, Gaia, ..

Le CDS n'est pas à l'origine un centre d'archivage, mais le devient par état de fait..

Aujourd'hui, c'est environ 1,000,000 requêtes/jour sur les services du CDS.

Présentation du CDS

Les service principaux du CDS:

- (1) Base de données d'objets astronomique : Simbad
 - → identification, bibliographie, données, mesures...
 - ~8,000,000 Objets



- (2) Collection de donnés issus des catalogues : VizieR
 - → issues de journaux, de logs d'observation ou de grand relevés principalement constitué de données tabulaires (mais aussi : spectres, images, série temporelles)
 - ~ 13,000 catalogues, 26,000 tables
- (3) Atlas interactif du ciel : **Aladin**
 - → découverte, visualisation et manipulation de données
 Base de données et archives d'images locales et distantes



Le rôle de l'Observatoire Virtuel

Le rôle de l'Observatoire virtuel (VO) dans la dissémination des données au CDS

- L'IVOA est une « alliance » internationale qui coordonne les moyens à mettre en œuvre pour rendre interopérable les ressources des centres de données en Astronomie.
- Standard sur les formats utilisés en plus des standards disciplinaire (FITS) et la description des données
- Standards sur les protocoles de dissémination

http://www.ivoa.net/

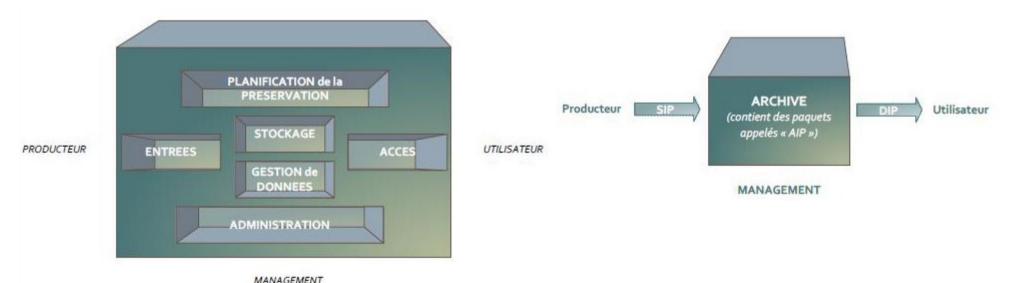


Le **D**ata **S**eal of **A**pproval (DSA) est obtenu en août 2014. Il s'agit d'une labellisation internationale pour la préservation des données scientifiques sur le long terme.

- La labellisation DSA nécessite une réévaluation tous les 2 ans
- Contient 16 critères, pouvant être résumé par :
 - Les données sont accessibles sur internet
 - Les licences/droits d'utilisations des données sont clairement exprimés
 - Les données sont réutilisable
 - Les données sont fiables
 - Les données sont identifiées de manière unique selon une nomenclature permettant un référencement externe
- Exigence d'une organisation de l'infrastructure selon un modèle reconnu internationalement et adaptée au cas spécifique
 - → OAIS -like

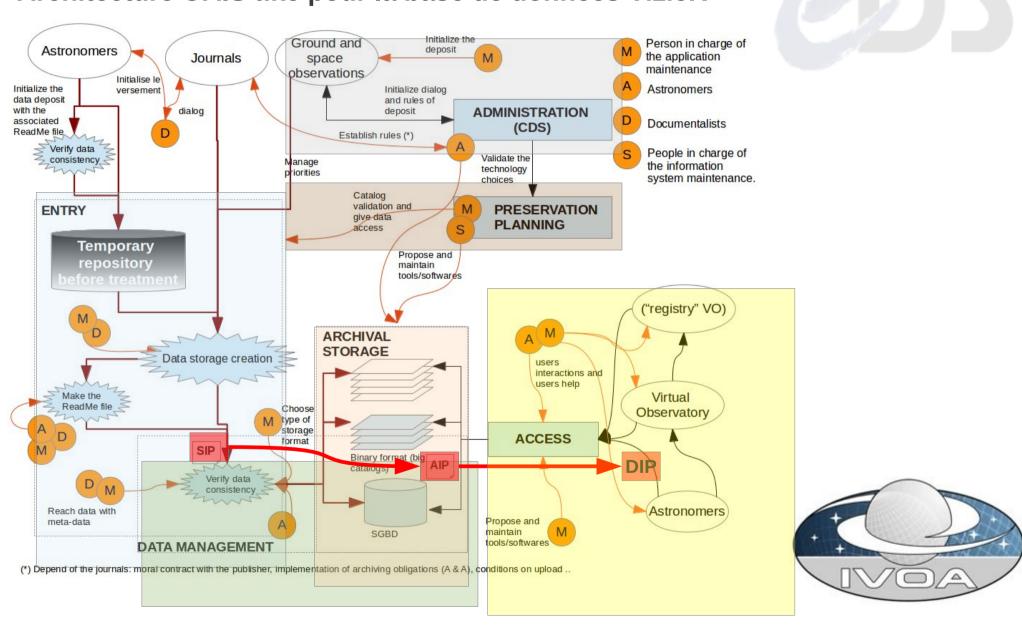
OAIS = Open Archive Information System (ISO 14721) (Système ouvert d'archivage d'information) est un modèle conceptuel destiné à la gestion, à l'archivage et à la préservation à long terme de documents numériques.

- Définit les acteurs/responsabilités dans le SI : producteur/utilisateur/manager
- Définit un modèle fonctionnel de l'archive en 6 composantes
- Définit les flux d'informations
 - → les paquets OAIS : SIP (en entrée), AIP (archives), DIP (diffusion)
- Définit des normes/recommandations « ouvertes » exemple : modalité de versements des données



Cines: https://www.cines.fr/archivage/un-concept-des-problematiques/le-modele-de-reference-loais/
https://www.cines.fr/archivage/un-concept-de-reference-loais/
https://www.cines.fr/archivage/un-concept-de-reference-loais/</

Architecture OAIS-like pour la base de données VizieR

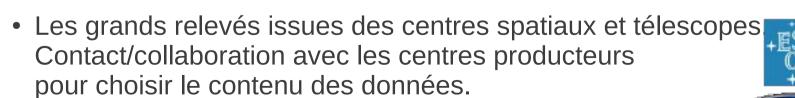


Pipeline d'ingestion des données pour le service VizieR

Les règles d'ingestion sont définies par le CDS (exprimées par le CDS sous la forme de règles/normes dans le modèle OAIS)

Des pipelines dédiés aux journaux et aux grands relevés :

- Données issues du journal A&A (contrat).
 Application web d'ingestion à l'usage des auteurs incluant la documentation et la liste des formats acceptés
- Données issues des autres journaux :
 AAS, MNRAS, etc. (accords tacites)
 Fournies par l'éditeur ou téléchargé depuis les sites web des journaux et mise en forme par les documentalistes du CDS









Pipeline d'ingestion des données issus des journaux

J/A+A/424/545

Optically faint obscured quasars

Discovery of optically faint obscured quasars with Virtual Observatory tools.
Padovani P., Allen M.G., Rosati P., Walton N.A.

ASTRON. ASTRONYS., 424, 545-559 (2004)>

=2004A&A...424..545P

ADC_Keywords: QSOs ; Active gal. nuclei ; X-ray sources
Keywords: astronomical data bases: miscellaneous - methods: statistical - qalaxies: 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 table2.dat table4.dat	90 90 90		Type 2 AGN candidates, HDF-N Type 2 AGN candidates, CDF-S Type 2 AGN candidates, UDF					

See also:

Byte-by-byte Description of file: table*.dat

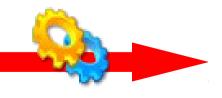
_	Bytes	Format	Units	Label	Explanations
	1- 19	A19		GOODS	GOODS designation (JHHMMSS.ss+DDMMSS.s)
	22- 25			UDF	? UDF designation (Cat. II/258, table 4 only)
	27- 29	13		A03	Alexander et al. (2003, Cat. <j 126="" 539="" aj=""> sequential number, [ABB2003] CDFN NNN (table1) or [ABB2003] CDFS NNN (table2&3) in Simbad</j>
	31- 33	13		S04	<pre>? Szokoly et al. (2004, Cat. <j 155="" 271="" apjs=""> sequential number, [SBH2004] XID NNNa in Simbad (table2 only)</j></pre>
	34	A1	m	S04	[a] Multiplicity index on SO4
	35- 36		h	RAh	Right ascension (J2000.0)
	38- 39	12	min	RAm	Right ascension (J2000.0)
	41- 45	F5.2	S	RAs	Right ascension (J2000.0)

Les données tabulaires sont accompagnées d'un fichier ReadMe décrivant les données

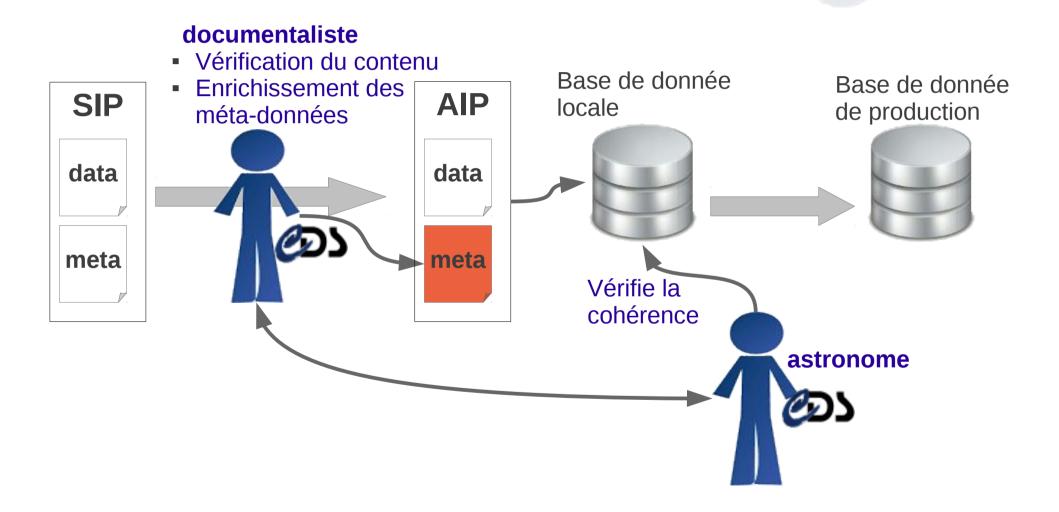
Le fichier (ASCII) constitue le paquet SIP ; il peut être construit par les éditeurs, les auteurs ou les documentalistes

Le fichier ReadMe est accompagné d'un fichier de configuration construit par les documentalistes pour enrichir les données/méta-données

Les données/méta-données tabulaires sont stockés sous forme de tables dans le SGBD



La fiabilisation des données en entrées

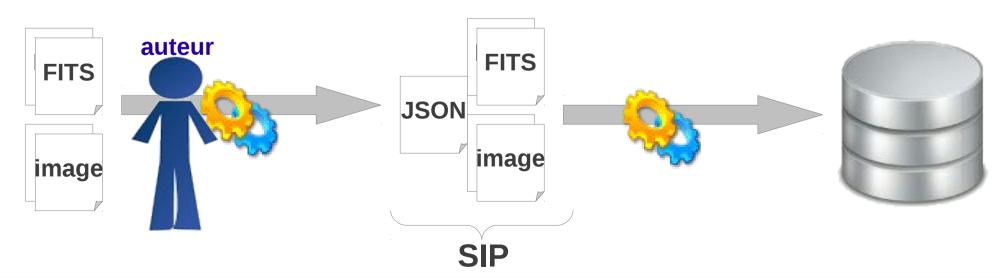


Ingestion de grandes volumétries dans VizieR

- → Les grands catalogues issues des centres spatiaux et télescopes
 - ~30 grands catalogues (parmi 13,000) stockés sur ~6Tb.
 souvent sur la totalité du ciel. Il peut comporter des images ou des données tabulaires
 La plus grande table Gaia (simulation Gaia) ~2milliards de lignes, ~40colonnes
 - Les catalogues les plus interrogés
 - Utilisation de formats adaptés pour gérer des données
 - 1) stockage selon des formats binaires propriétaire au CDS et sont interrogeable par le service VizieR avec des formats en sortie : HTML, TSV, VOTable (XML)
 - 2) stockage en SGBD (pour les besoins de l'Observatoire Virtuel)

Ingestion des données non tabulaires dans VizieR (en développement)

- → Images, spectres, séries temporelles.. issues des articles
 - Aujourd'hui archivés au CDS et disponible par FTP ou par des liens dans les pages de résultats de VizieR
 - Indexation et méta-données en cours de développement
 - → discussions avec les journaux (AAS) pour définir le format des méta-donnés
 - Demande aux auteurs leur contribution pour ajouter les méta-données



La planification et la préservation des archives

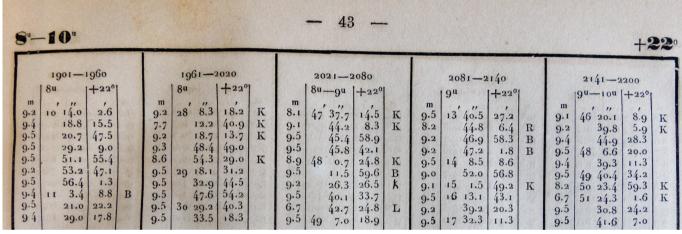
La préservation des archives numérique concerne l'ensemble du SI

- 1) L'ensemble des logiciels pour l'ingestion et la dissémination des données
- 2) Le format de stockage des données adapté : SGBD (méta-données et tables de taille modeste), format binaires (pour les grandes volumétries), etc
- 3) L'infrastructure informatique constituée des serveurs, disques, etc.
 - → réplication, miroirs, etc

Les migrations / évolutions du CDS

Le catalogue Bonner Durchmusterung (Argelander 1859-1903)

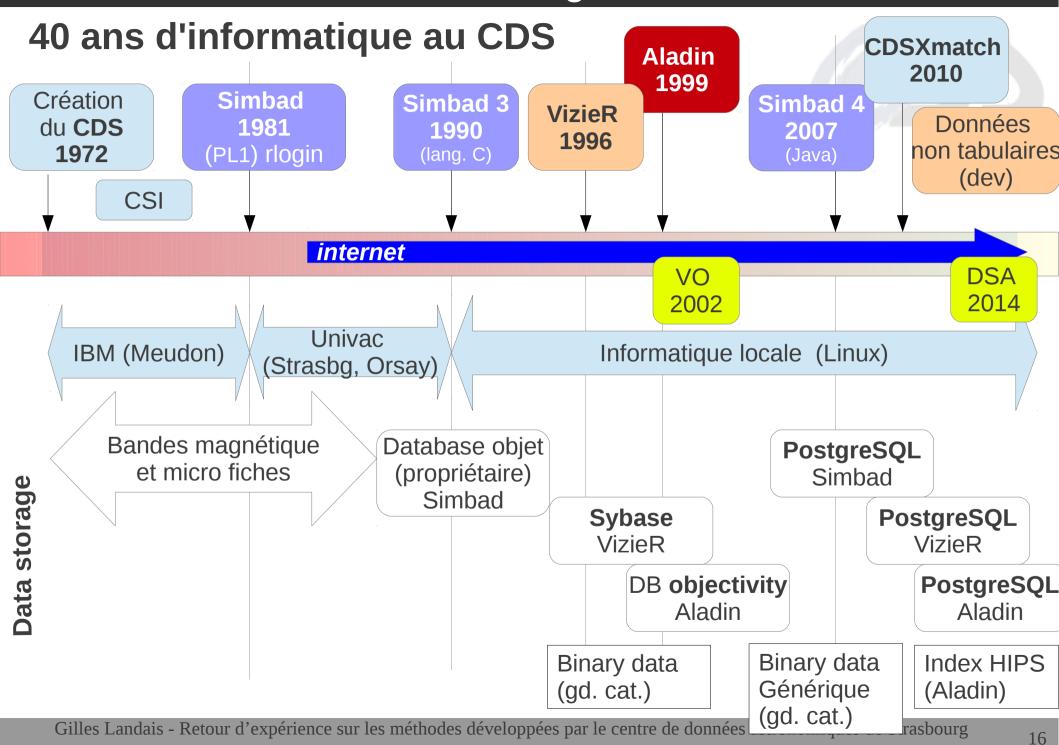






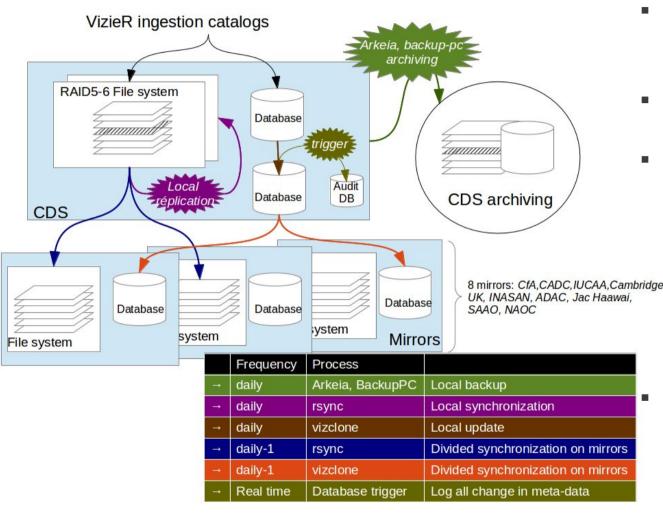
Full	_RAJ2000	_DEJ2000	zonesign	zone	<u>num</u>	suppl	mag	RA1855	DE1855	_RA.icrs	_DE.icrs
	"h:m:s"	"d:m:s"		deg			mag	"h:m:s"	"d:m:s"	"h:m:s"	<u>"d:m:s"</u>
_ △▼	Δ▼	Δ▼	△▼	△▼	ΔΨ	ΔΨ	ΔΨ	△▼	ΔΨ	Δ▼	△▼
1	08 18 44.2	+21 35 46	+	22	1901		9.2	08 10 14.0	+22 02.6	08 18 44.2	+21 35 46
2	08 18 49.7	+21 48 40	+	22	1902		9.4	08 10 18.8	+22 15.5	08 18 49.7	+21 48 40
3	08 18 53.4	+22 20 39	+	22	1903		9.5	08 10 20.7	+22 47.5	08 18 53.4	+22 20 39
4	08 18 59.7	+21 42 08	+	22	1904		9.5	08 10 29.2	+22 09.0	08 18 59.7	+21 42 08
<u>5</u>	08 19 24.1	+22 28 28	+	22	1905		9.5	08 10 51.1	+22 55.4	08 19 24.1	+22 28 28
<u>6</u>	08 19 25.7	+22 20 09	+	22	1906		9.2	08 10 53.2	+22 47.1	08 19 25.7	+22 20 09

Les migrations / évolutions du CDS



Archivage et préservation au CDS

Réplication des données



- Localement (dans un autre bâtiment)
- Backup
- Utilisation de sites miroirs : Simbad : CDS+CfA (USA)

VizieR: 8 miroirs

Aladin: 1 miroir à l'IAS(Orsay)

- → génère une maintenance significative
- Utilisation d'un système de balancing en cas d'indisponibilité au CDS (GLU)

Dissémination des données

Des données en libre accès

- Des applications web ou standalone (Aladin) dédiées à chaque service
- Accès via les outils de l'Observatoire Virtuel
- Accès par scripts
- Les résultats des requêtes sont disponibles dans des formats reconnus par la communauté astronomique: TSV, VOTable, FITS
- Des recherches de ressources (catalogue/table/images...) facilitées par des méta-données permettant des recherches multicritères.
- Une indexation efficace

Une identification pérenne et reconnue par la communauté

- Un identifiant « bibcode » est attribué à chaque article
- Un identifiant est attribué à chaque catalogue

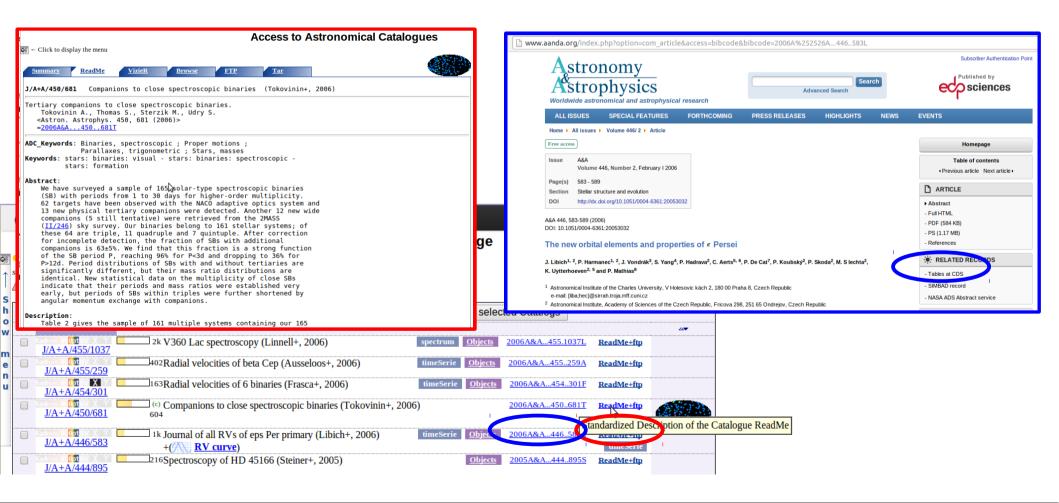
Le « bibcode » et l'identifiant catalogue sont reconnus et utilisés par l'application de bibilographie ADS qui indexe les articles parus en astronomie.

L' identifiant catalogue est utilisé dans le « registry » de l'Observatoire Virtuel pour être utilisé par des application externes.

Dissémination des données

L'accès aux méta-données

- Les méta-données sont distribués dans les résultats des requêtes (VOTable)
- Tous les catalogues contiennent un fichier ReadMe accessible via le site web : il contient notamment un lien vers l'article original : littérature, licences ...



Le rôle du VO dans la dissémination des données

La réutilisation des données est garantie par l'utilisation de standards

- En utilisant des sorties standards (VOTable/XML) comprise par des applications du VO comprenant les données+ les méta-données
- En utilisant une nomenclature standard pour désigner le contenu des colonnes des tables
- En implémentant les services VO pour accéder aux données :
 - Recherche spatiale
 - Interrogation des tables par un langage commun basé sur SQL et incluant des fonctionnalités d'astronomie

Une visibilité des données du CDS à travers l'Observatoire Virtuel Le registry VO est un « dictionnaire de ressources », basé sur le standard OAI-PMH, est interrogé par les outils VO

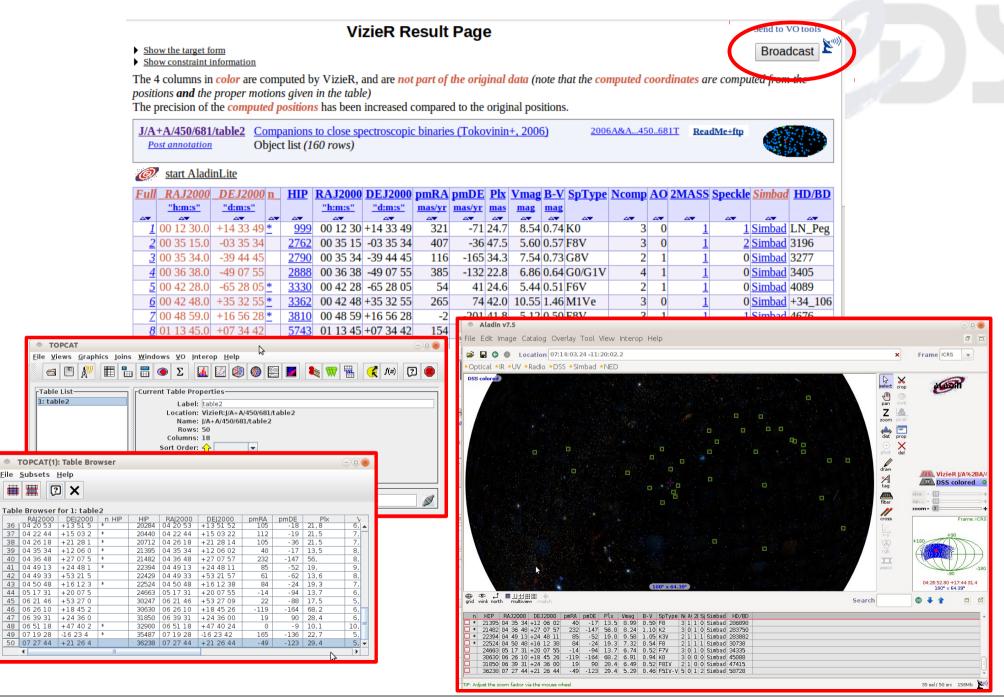
Le rôle du VO dans la dissémination des données

Exemple de standardisation : le format de sortie VOTable

Les méta-données sont distribués dans les résultats des requêtes (VOTable)

```
vizier votable.vot (~/Téléchargements) - GVIM1
                                                                                                                                 <?xml version="1.0" encoding="UTF-8"?>
<VOTABLE version="1.2" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"</pre>
  xmlns="http://www.ivoa.net/xml/V0Table/v1.2"
  xsi:schemaLocation="http://www.ivoa.net/xml/VOTable/v1.2 http://www.ivoa.net/xml/VOTable/v1.2">
 <DESCRIPTION>
  VizieR Astronomical Server vizier.china-vo.org
 </DESCRIPTION>
<RESOURCE ID="yCat 1239" name="I/239">
  <DESCRIPTION>The Hipparcos and Tycho Catalogues (ESA 1997)/DESCRIPTION>
  <C00SYS ID="J2000 1991.250" system="eq FK5" equinox="J2000" epoch="1991.250"/>
  <TABLE ID="I 239 hip main" name="I/239/hip main">
   <DESCRIPTION>The Hipparcos Main Catalogue
   <FIELD name="HIP" ucd="meta.id;meta.main" datatype="int" width="6">
     <DESCRIPTION>Identifier (HIP number) (H1)/DESCRIPTION>
   </FIELD>
    <FIELD name="RA(ICRS)" ucd="pos.eq.ra;meta.main" datatype="double" width="12" precision="8" unit="deg">
     <DESCRIPTION>*? alpha, degrees (ICRS, Epoch=J1991.25) (H8)/DESCRIPTION>
     <VALUES null="NaN" />
   </FIELD>
   <FIELD name="DE(ICRS)" ucd="pos.eg.dec;meta.main" datatype="double" width="12" precision="8" unit="deg">
     <DESCRIPTION>*? delta, degrees (ICRS, Epoch=J1991.25) (H9)</DESCRIPTION>
     <VALUES null="NaN" />
   </FIELD>
<DATA><TABLEDATA>
<TR><TD>1</TD><TD>0.00091185</TD><TD>1.08901332</TD></TR>
<TR><TD>2</TD><TD>0.00379737</TD><-TD>-19.49883745</TD></TR>
<TR><TD>3</TD>0.00500795</TD>38.85928608</TD></TR>
<TR><TD>4</TD><TD>0.00838170</TD><TD>-51.89354612</TD></TR>
<TR><TD>5</TD><TD>0.00996534</TD><-40.59122440</TD></TR>
<TR><TD>6</TD><TD>0.01814144</TD><TD>3.94648893</TD></TR>
<TR><TD>7</TD><TD>0.02254891</TD><TD>20.03660216</TD></TR>
<TR><TD>8</TD><TD>0.02729160</TD><TD>25.88647445</TD></TR>
<TR><TD>9</TD><TD>0.03534189</TD><TD>36.58593777</TD></TR>
<TR><TD>10</TD><TD>0.03625309</TD><TD>-50.86707360</TD></TR>
<TR><TD>11</TD><TD>0.03729695</TD><TD>46.94000154</TD></TR>
<TR><TD>12</TD><TD>0.04091756</TD><TD>-35.96022482</TD></TR>
<TR><TD>13</TD><TD>0.04167970</TD><TD>-22.59468060</TD></TR>
<TR><TD>14</TD><TD>0.04827189</TD><TD>-0.36042119</TD></TR>
<TR><TD>15</TD><TD>0.05030890</TD><TD>50.79117384</TD></TR>
                                                                                                                                                                 Haut
```

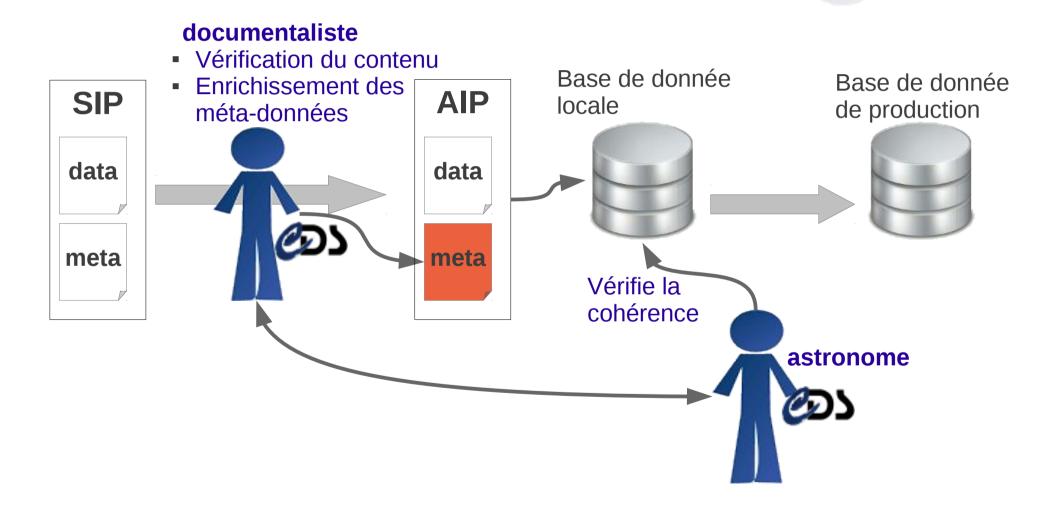
Le rôle du VO dans la dissémination des données



Conclusions

- Difficultés rencontrés pour la labellisation DSA : contrats, bon usage des données, licences...
- Evolution des formats de stockage en entrée sous des formats structurés : VOTABLE, JSON
- Modèle OAIS-like
 - Un cadre adapté à la structure et à l'architecture du CDS
 - Clarification de certaines questions comme les modalités d'ingestion de données, les licences...
 - Importance d'une « archive ouverte» tant pour les données que les méta-données
- Être présent dans les centres de décisions
 - Avec les centres producteurs : journaux, grand centre spatiaux (ESO, ESA, NASA ...)
 - Avec des partenaires (internationaux) : Observatoire Virtuel, ADS...
- Être proche des utilisateurs : dans des meetings (AAS), ...

La fiabilisation des données en entrées



Les grands partenaires du CDS

Les relations du CDS avec les centres producteur de données

- Relation avec les grands producteur de données (ESO, ESA, NASA, ...)
- Les accords avec les journaux par contrat (A&A)
- Discutions avec les journaux : (ex : AAS pour le choix des méta-données et pour la mise en place des pipelines)

Les relations du CDS avec des partenaires internationaux

- Le CDS est un acteur clé dans les prises de décisions de l'Observatoire Virtuel: propositions,...
- Des relations avec des partenaires comme celle avec le portail web pour la bibliographie en astronomie (ADS).

Relation avec les astronomes

- Les astronomes du CDS forment le 1/3 de l'équipe du CDS
- Les réunions annuelles de l'AAS permet de rencontrer les astronomes américains

Le conseil scientifique

Il est composé de 6 membres français et 6 membres étranger. Il évalue de manière critique le CDS : émet des propositions , ...