Description Sémantique par les "Unified Content Descriptors"

Sébastien DERRIERE – derriere@astro.u-strasbg.fr
Centre de Données astronomiques de Strasbourg (CDS)

Plan

- Contexte: (méta)données en astronomie
- Définition des UCD
- Applications
- Perspectives

Contexte: (méta)données en astronomie

- Les astronomes font face à de grands volumes de données hétérogènes:
  - articles, publications
  - catalogues (données tabulaires)
  - images, relevés du ciel
  - spectres, cubes de données
- Données produites par de nombreux projets: comment les comparer?

Tables astronomiques

- Service Vizier au CDS: 4000 catalogues différents
  - 1 catalogue peut contenir plusieurs tables
  - 1 table peut avoir + de 100 colonnes
  - 1 catalogue = jusqu'à un milliard d'objets
- Origines variées pour les catalogues:
  - publication électronique de journaux
  - numérisation de plaques photos
  - relevés du ciel avec CCD

Description des catalogues astronomiques (tables):

214.386166	-57.767818	16.926	15.777	99.999	0.09	0.19	9.99	17.067	15.508	99.999
214.535889	-57.767764	16.458	15.562	99.999	0.07	0.17	9.99	16.496	15.457	99.999
214.401036	-57.767685	14.974	14.391	99.999	0.04	0.11	9.99	15.021	14.549	99.999
214.569711	-57.767623	17.971	15.777	99.999	0.18	0.19	9.99	17.394	15.553	99.999
214.349915	-57.767576	16.975	99.999	99.999	0.10	9.99	9.99	16.840	99.999	99.999
214.550993	-57.767487	16.801	15.716	99.999	0.09	0.18	9.99	16.605	15.682	99.999
214.557370	-57.767406	99.999	16.525	13.594	9.99	0.27	0.22	99.999	15.544	12.905
214.404212	-57.767370	15.848	14.973	99.999	0.05	0.13	9.99	15.654	15.197	99.999
214.296113	-57.767262	15.161	13.266	99.999	0.04	0.08	9.99	15.055	13.271	99.999
214.238914	-57.767254	15.363	14.061	99.999	0.04	0.10	9.99	14.916	14.106	99.999
214.286765	-57.767228	15.694	13.984	99.999	0.05	0.09	9.99	15.784	14.019	99.999
214.595510	-57.767131	17.716	16.170	99.999	0.15	0.23	9.99	17.274	16.150	99.999
214.466317	-57.767040	15.975	13.680	12.353	0.06	0.09	0.12	15.998	13.686	12.836
214.503014	-57.767008	17.436	99.999	99.999	0.13	9.99	9.99	17.548	99.999	99.999
214.483010	-57.766971	99.999	16.015	99.999	9.99	0.21	9.99	99.999	16.370	99.999
214.47[illegible]701	-57.[illegible]66933	16.[illegible]31	13.[illegible]17	12.[illegible]08	0.06	0.09	0.14	16.025	13.909	13.395
214.38[illegible]021	-57.[illegible]66657	18.[illegible]85	99.[illegible]99	99.[illegible]99	0.19	9.99	9.99	19.044	99.999	99.999
214.26[illegible]358	-57.[illegible]66521	17.[illegible]67	15.[illegible]84	99.[illegible]99	0.11	0.20	9.99	17.209	16.248	99.999
214.61[illegible]797	-57.[illegible]66361	17.[illegible]03	15.[illegible]49	99.[illegible]99	0.10	0.14	9.99	16.741	14.812	99.999
214.26[illegible]771	-57.[illegible]66321	17.[illegible]88	15.[illegible]12	99.[illegible]99	0.13	0.18	9.99	16.713	15.023	99.999
214.53[illegible]274	-57.[illegible]66314	16.[illegible]79	14.[illegible]23	13.[illegible]36	0.06	0.11	0.16	16.099	14.255	13.102
214.56[illegible]229	-57.[illegible]66304	17.[illegible]49	15.[illegible]55	99.[illegible]99	0.11	0.20	9.99	17.213	17.242	99.999
214.25[illegible]734	-57.[illegible]66279	16.[illegible]55	14.[illegible]60	13.[illegible]44	0.07	0.12	0.18	16.426	14.683	12.860
214.49[illegible]574	-57.[illegible]66278	17.[illegible]09	14.[illegible]56	13.[illegible]33	0.10	0.13	0.22	16.628	14.731	13.337
214.59[illegible]738	-57.[illegible]66200	17.[illegible]74	99.[illegible]99	99.[illegible]99	0.11	9.99	9.99	16.903	99.999	99.999
214.31[illegible]793	-57.[illegible]66161	15.[illegible]16	14.[illegible]37	99.[illegible]99	0.05	0.12	9.99	15.496	14.697	99.999
214.34[illegible]976	-57.[illegible]66060	99.[illegible]99	15.[illegible]04	99.[illegible]99	9.99	0.19	9.99	99.999	15.705	99.999
214.27[illegible]051	-57.[illegible]66037	16.[illegible]67	15.[illegible]25	99.[illegible]99	0.10	0.18	9.99	17.014	16.130	99.999
214.58[illegible]047	-57.[illegible]65921	15.[illegible]41	13.[illegible]37	12.[illegible]39	0.06	0.09	0.12	15.858	13.620	12.295
214.42[illegible]628	-57.[illegible]65900	99.[illegible]99	16.[illegible]14	13.[illegible]28	9.99	0.23	0.27	99.999	16.593	13.955
214.41[illegible]201	-57[illegible]65870	15[illegible]41	14[illegible]98	99[illegible]99	0.05	0.12	9.99	15.604	14.415	99.999
214.517290	-57.765842	99.999	16.320	99.999	9.99	0.24	9.99	99.999	18.163	99.999
214.4[illegible]607	-5[illegible]65814	[illegible]	[illegible]	[illegible]	0.17	0.19	9.99	17.819	15.194	99.999
214.554082	-57.765781	15.537	12.979	11.402	0.05	0.08	0.09	15.540	12.948	11.385
214.474225	-57.765746	18.131	99.999	99.999	0.20	9.99	9.99	18.831	99.999	99.999

RA DEC magI magJ magK

Description des tables

- Une description standardisée (ReadMe) est associée à chacun des 4000 catalogues de VizieR... mais:
  - Les origines étant hétérogènes, les descriptions le sont aussi
  - Chaque auteur de catalogue peut nommer les colonnes à sa guise:
    - ex: Magnitude V = Vmag, V, mV, V1, V2, ... plus de 140 noms de colonnes pour une même quantité !!
- A priori, pas de comparaison automatisée possible entre différentes tables

Description des tables

- Comment rechercher des catalogues d'après leur contenu?
  - catalogues contenant une mesure de mouvement propre
  - catalogues contenant une magnitude Johnson B
- Comment identifier les quantités comparables dans différents catalogues?

Les UCD

- UCD = Unified Content Descriptors
- Objectif: donner une description sémantique du contenu des colonnes
- UCD1: première version (ESO-CDS data mining project)
  - Exploration manuelle de 100,000 colonnes de VizieR
  - Création de mots pour chaque quantité trouvée
  - Définition de ~1500 termes (UCD1) pour la description sémantique du contenu de VizieR
    - ex: PHOT_EXTINCTION_ISM = Interstellar extinction, POS_GAL_LAT = Galactic Latitude

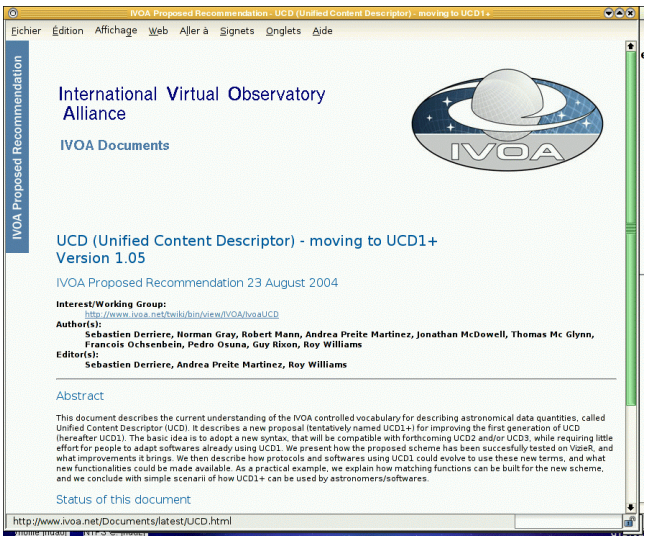
L'Observatoire Virtuel (VO)

- Nombreux projets (inter)nationaux d'Observatoire Virtuel Astronomique
- Coordination / coopération:
  - IVOA (International Virtual Observatory Alliance)
  - Définition de standards communs pour faciliter l'interopérabilité
- Lancement du VO après la définition des UCD1

UCD1+

- Migration vers les UCD1+ avec le VO
  - Réutiliser une base de connaissance existante (UCD1)
  - Définition plus souple des termes
  - Vocabulaire contrôlé par l'IVOA
- Liste de mots standards qui peuvent être combinés pour exprimer la sémantique:
  - phot.mag;em.opt.V
  - stat.error;phys.temperature
  - pos.eq.dec

UCD1+ et IVOA



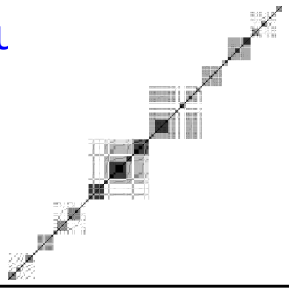
Processus de validation et documentation au niveau de l'IVOA

UCD1+

- Même objectif que les UCD1:
  - décrire des quantités ("qu'est-ce que c'est?")
  - avec un niveau de détail raisonnable
  - pour faire des comparaisons entre ensembles hétérogènes (interopérabilité)
- Changement de syntaxe:
  - mots composés d'atomes (pos.gal.lat)
  - possibilité de combiner plusieurs mots avec des ';' (phys.temperature;instr.tel)
- Moins de 500 mots différents dans le vocabulaire

UCD1+

- Plus de souplesse et de précision
- On peut réutiliser un mot dans plusieurs UCD
  - phot.mag;em.opt.B
  - phot.mag;em.opt.R
- Le premier mot porte l'essentiel du sens
- Possibilité de définir des fonctions de comparaisons "floues" entre deu



Applications et outils

- Utilisation dans les standards et protocoles du VO:
  - VOTable (format XML d'échange de tables)

Exemple de VOTable 1.1:

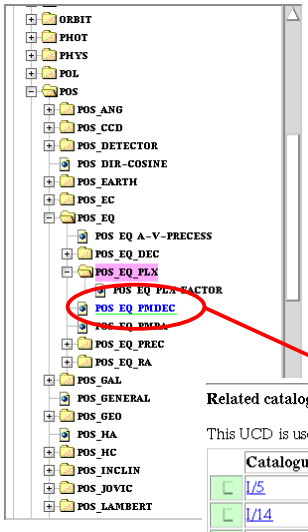
METADATA

DATA

```
<?xml version="1.0"?>
<VOTABLE version="1.1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/VOTable/VOTable/v1.1">
  <COOSYS ID="J2000" equinox="J2000." epoch="J2000." system="eq_FK5"/>
  <RESOURCE name="myFavouriteGalaxies">
    <TABLE name="results">
      <DESCRIPTION>Velocities and Distance estimations</DESCRIPTION>
      <PARAM name="Telescope" datatype="float" ucd="phys.size;instr.tel"
             unit="m" value="3.6"/>
      <FIELD name="RA"   ID="col1" ucd="pos.eq.ra;meta.main" ref="J2000"
             datatype="float" width="6" precision="2" unit="deg"/>
      <FIELD name="Dec"  ID="col2" ucd="pos.eq.dec;meta.main" ref="J2000"
             datatype="float" width="6" precision="2" unit="deg"/>
      <FIELD name="Name" ID="col3" ucd="meta.id;meta.main"
             datatype="char" arraysize="8*"/>
      <FIELD name="RVel" ID="col4" ucd="src.veloc.hc" datatype="int"
             width="5" unit="km/s"/>
      <FIELD name="e_RVel" ID="col5" ucd="stat.error;src.veloc.hc"
             datatype="int" width="3" unit="km/s"/>
      <FIELD name="R" ID="col6" ucd="phys.distance" datatype="float"
             width="4" precision="1" unit="Mpc">
        <DESCRIPTION>Distance of Galaxy, assuming H=75km/s/Mpc</DESCRIPTION>
      </FIELD>
      <DATA>
        <TABLEDATA>
        <TR>
          <TD>010.68</TD><TD>+41.27</TD><TD>N  224</TD><TD>-297</TD><TD>5</TD><TD>0.7</TD>
        </TR>
        <TR>
          <TD>287.43</TD><TD>-63.85</TD><TD>N 6744</TD><TD>839</TD><TD>6</TD><TD>10.4</TD>
        </TR>
        <TR>
          <TD>023.48</TD><TD>+30.66</TD><TD>N  598</TD><TD>-182</TD><TD>3</TD><TD>0.7</TD>
        </TR>
        </TABLEDATA>
      </DATA>
    </TABLE>
  </RESOURCE>
</VOTABLE>
```


Applications et outils

- Utilisation dans les standards et protocoles du VO:
  - VOTable (format XML d'échange de tables)
  - Registry (description des catalogues dans les annuaires de resources)
  - ConeSearch (protocole d'interrogation des catalogues)
- Trouver des catalogues à partir de leur contenu



Trouver des catalogues intéressants
à partir des UCDs

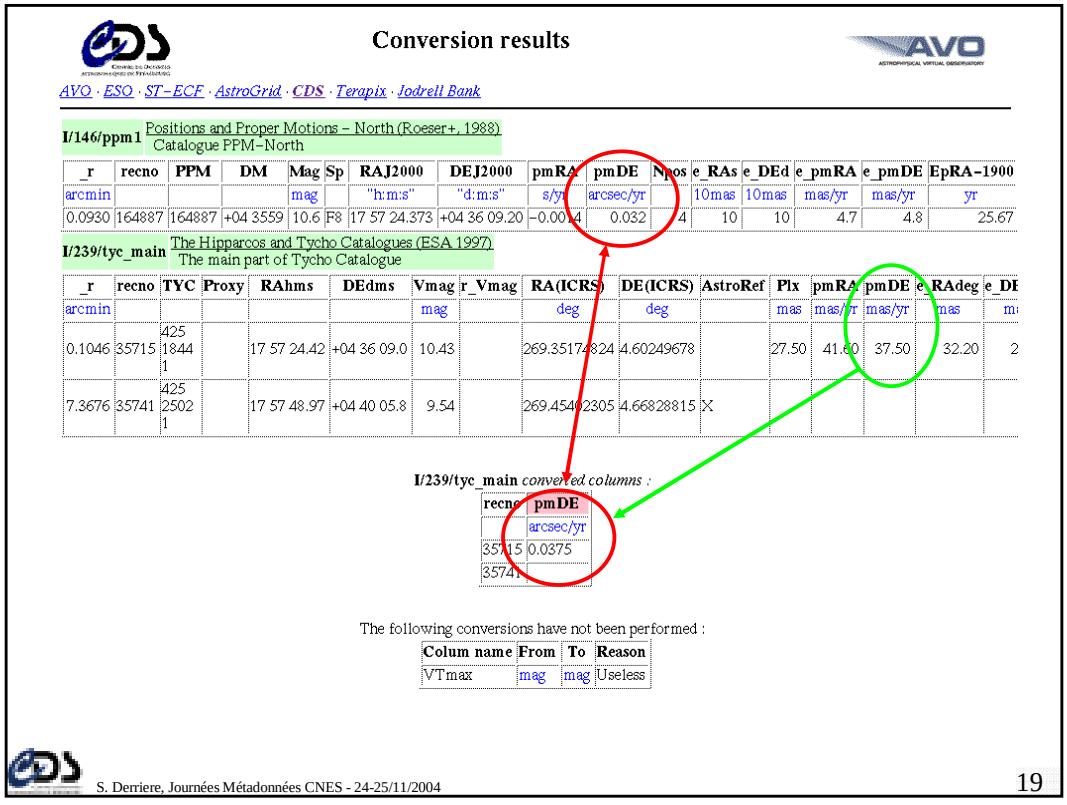
Related catalogues in VizieR:

This UCD is used in 209 columns, in 134 different catalogues (187 tables) of VizieR.

	Catalogue	Title	Bibcode
	I/5	Proper Motions in Cape Zone Catalogue -40/-52 (Spencer Jones H.+ 1936)	
	I/14	Proper Motions of 1160 Late-Type Stars (Fogh Olsen, 1970)	1970A&AS....2...69O
	I/40	WASHINGTON 20 Catalog (Morgan, 1933)	
	I/61B	AGK3 Catalogue (Dieckvoss, Heckmann 1975)	1975QB6.A15.....D
	I/62C	Perth 70: Positions of 24900 Stars (Hog+ 1976)	1976AAHam...9....1H
	I/68A	Positions and Proper Motions in alpha Per cluster (Fresneau, 1980)	1980BICDS..18...81F

Applications et outils

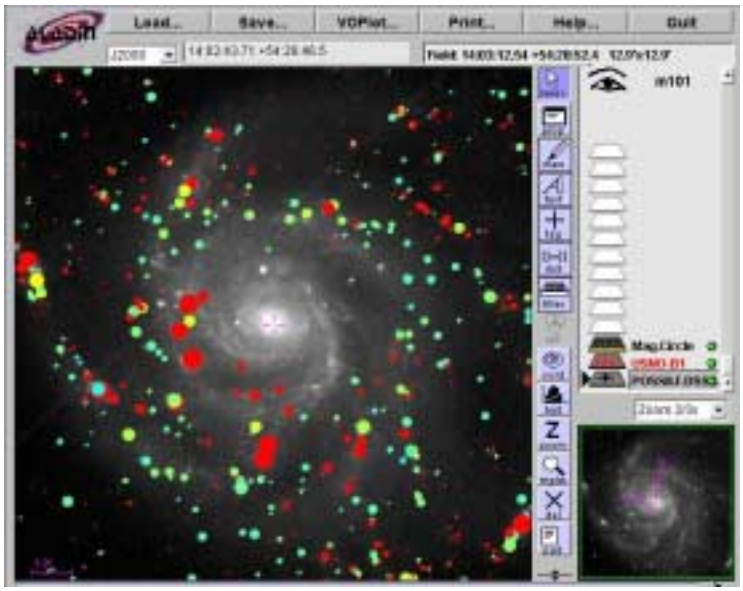
- Utilisation dans les standards et protocoles du VO:
  - VOTable (format XML d'échange de tables)
  - Registry (description des catalogues dans les annuaires de resources)
  - ConeSearch (protocole d'interrogation des catalogues)
- Trouver des catalogues à partir de leur contenu
- Comparaisons automatiques:
  - Si deux colonnes sont décrites par le même UCD, on peut les comparer



Intégration dans des outils

- L'utilisation des UCD dans VOTable permet d'introduire dans les applications des traitements génériques pour les catalogues
  - Une application de projection pourra par exemple rechercher les colonnes avec les UCD pos.eq.ra et pos.eq.dec pour superposer des objets sur une image
  - On peut définir des fonctions assez génériques sans connaître la structure ni le contenu du catalogue
    - filtres dans Aladin

Filtres dans Aladin

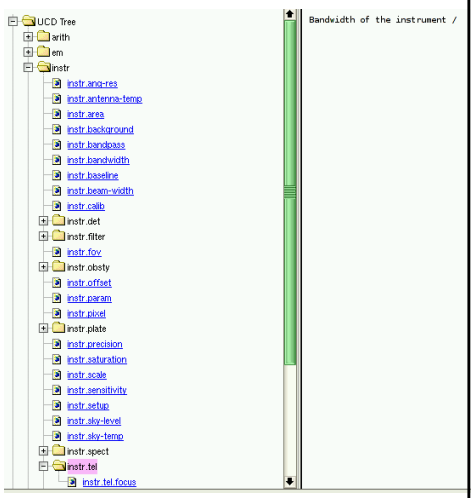


Mise en oeuvre

- Documents de référence:
  - http://vizier.u-strasbg.fr/UCD/
- Les fournisseurs de données ne sont pas contraints d'utiliser les UCD pour la gestion interne de leurs données:
  - utilisation d'une "translation layer" pour ajouter les UCD à un VOTable
  - UCD seulement nécessaire pour permettre l'intéropérabilité lors des échanges de données

Assignation d'UCD

- Trouver l'UCD pertinent pour décrire une quantité:
  - exploration de la liste de mots
  - recherche à partir d'une description textuelle (+ unité et nom de colonne)
- Post-doc INRIA (collab. LORIA)
  - Catégorisation textuelle
  - Algorithme bayesien, Rocchio



Assignation d'UCD

UCD	Dataset	Type	Name	Unit	Description
ID_MAIN PHYS_ABUND_[FE/H] PHOT_JHN_H ID_CATALOG PHOT_FLUX_HALPHA	hip_main.dat	A	Catalog	---	Catalogue (H=Hipparcos) (H0)
ID_NUMBER	hip_main.dat	I	HIP	---	Identifier (HIP number) (H1)
REMARKS CODE_MISC	hip_main.dat	A	Proxy	---	Note on Proxy: this flag provides a coarse indication of the presenceof nearby objects within 10arcsec of the given entry. If non-blank, it indicates that 'H' there is one or more distinct Hipparcos Catalogue entries, or distinct components of system from h_dm_com.dat 'T' there is one or more Proximity flag (H2)
POS_EQ_RA_MAIN	hip_main.dat	A	RAhms	---	Right ascension in h m s, ICRS (J1991.25) (H3)
POS_EQ_DEC_MAIN	hip_main.dat	A	DEdms	---	Declination in deg ' ", ICRS (J1991.25) (H4)
PHOT_JHN_V	hip_main.dat	F	Vmag	mag	? Magnitude in Johnson V (H5)
CODE_VARIAB	hip_main.dat	I	VarFlag	---	Note on VarFlag: the values are1: < 0.06mag ; 2: 0.06-0.6mag ; 3: >0.6mag ? Coarse variability flag (H6)
REFER_CODE	hip_main.dat	A.@c	r_Vmag	---	Note on r_Vmag: the source isG = ground-based, H=HIP, T=Tycho Source of magnitude (H7)

Perspectives: ontologies

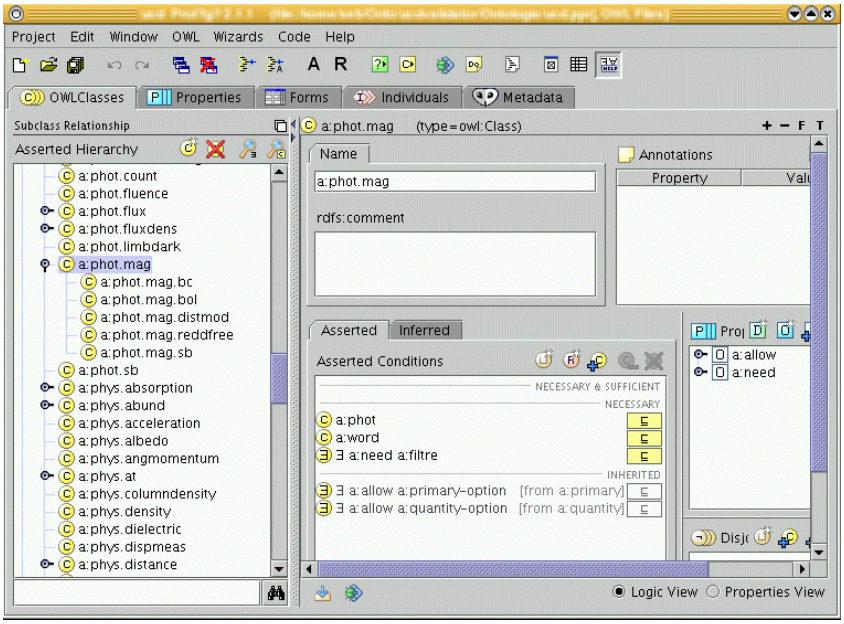
- Une ontologie définit un vocabulaire commun pour les personnes voulant partager de l'information dans un domaine
- Elle inclut des définitions de concepts de base du domaine et des relations entre eux
- Interprétable par machines et humains
- Objectifs pour en construire une:
  - Décrire (base de connaissance)
  - Raisonner (inference)

Ontologies

- Lien évident avec les UCD et le thesaurus de l'UAI
- Relation avec le Semantic Web
- Stage (été 2004)
  - Editeur: Protégé
  - Sortie en OWL (Ontology Web Language)
  - Application aux UCD
  - A la limite des standards actuels (gestions de quantités numériques!)
- Poursuite avec le programme VO-Tech

Ontologies

Prototype pour la validation des UCD1+ avec une ontologie



Conclusions

- UCD = description sémantique simple
- Utilisés dans les standards du VO
- Permettent l'intéropérabilité des services
- Première étape dans la construction d'une ontologie de l'astronomie