From astronomical knowledge bases to astronomical ontologies

Andrea Preite-Martinez
Sebastien Derriere

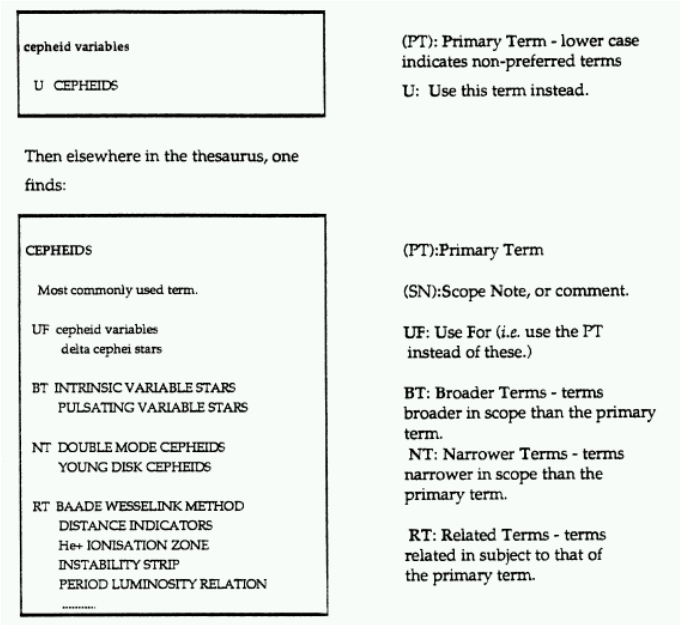
Astronomical Knowledge Bases

- Astronomers already have built various knowledge bases:
  - IAU thesaurus
  - UCDs
  - SIMBAD object types
  - Keywords from journals
- How to build ontologies on top of these ?
- Possible use cases

IAU thesaurus

- Compilation by astronomy librarians for on-line access to information
- Elaborated in the early 1990s
- Hierarchy of ~1500 terms
  - ~3300 expressions
  - covers various topics
- See Hernandez and Mothe's talk

IAU thesaurus



cepheid variables

U CEPHEIDS

(PT): Primary Term - lower case indicates non-preferred terms

U: Use this term instead.

Then elsewhere in the thesaurus, one finds:

CEPHEIDS

Most commonly used term.

UF cepheid variables
delta cephei stars

BT INTRINSIC VARIABLE STARS
PULSATING VARIABLE STARS

NT DOUBLE MODE CEPHEIDS
YOUNG DISK CEPHEIDS

RT BAADE WESSELINK METHOD
DISTANCE INDICATORS
He+ IONISATION ZONE
INSTABILITY STRIP
PERIOD LUMINOSITY RELATION
....

(PT):Primary Term

(SN):Scope Note, or comment.

UF: Use For (*i.e.* use the PT instead of these.)

BT: Broader Terms - terms broader in scope than the primary term.

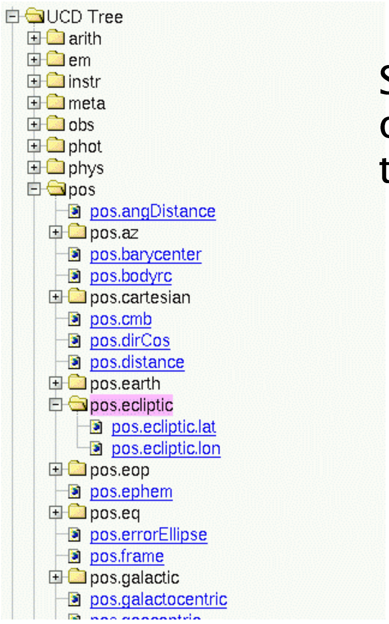
NT: Narrower Terms - terms narrower in scope than the primary term.

RT: Related Terms - terms related in subject to that of the primary term.

UCDs

- Description of astronomical quantities, measurements
- Hierarchical structure, ~450 words
- Originate from large collection of actual astronomical catalogue descriptions
- Revised by the IVOA for Virtual Observatories

UCDs



- UCD Tree
  - arith
  - em
  - instr
  - meta
  - obs
  - phot
  - phys
  - pos
    - pos.angDistance
    - pos.az
    - pos.barycenter
    - pos.bodyrc
    - pos.cartesian
    - pos.cmb
    - pos.dirCos
    - pos.distance
    - pos.earth
    - pos.ecliptic
      - pos.ecliptic.lat
      - pos.ecliptic.lon
    - pos.eop
    - pos.ephem
    - pos.eq
    - pos.errorEllipse
    - pos.frame
    - pos.galactic
    - pos.galactocentric
    - pos.geocentric

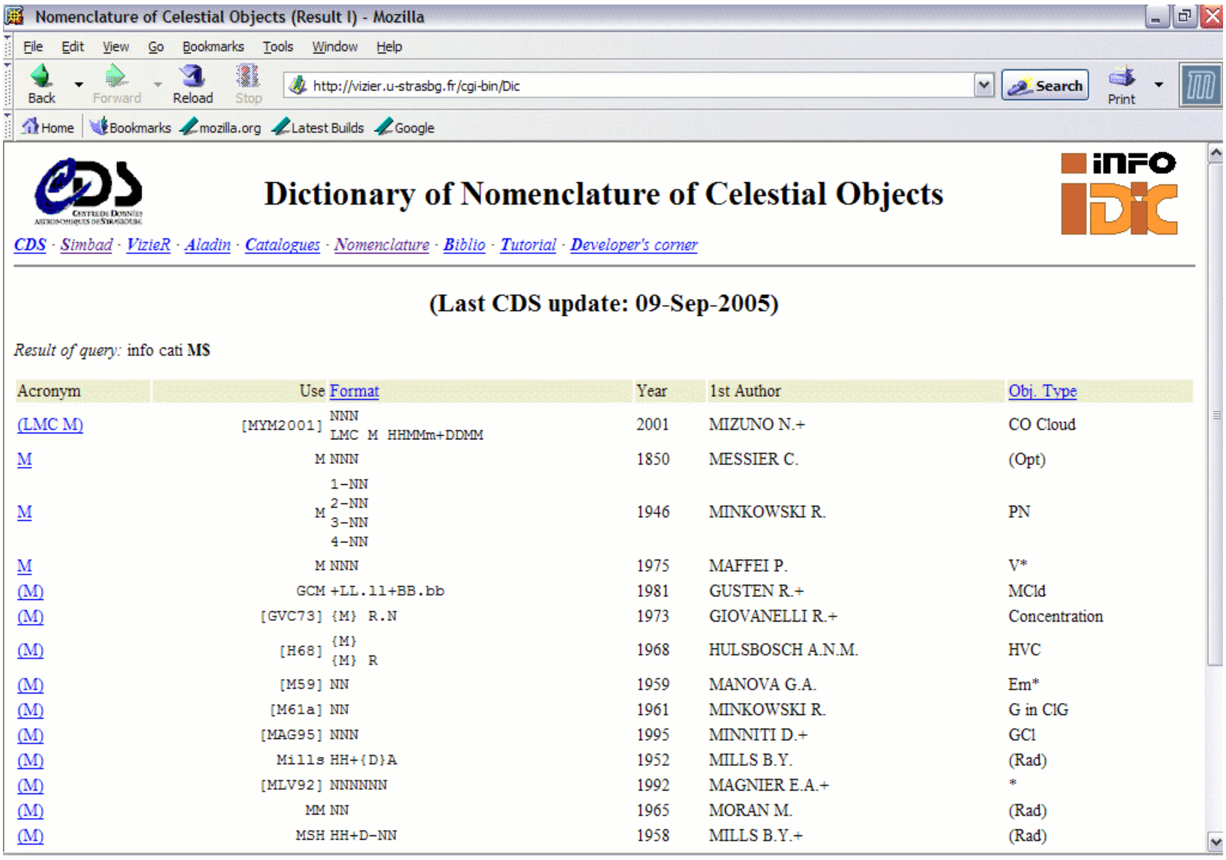
Standardized description of what is measured in tables.

SIMBAD object types

- Hierarchical classification of astronomical objects in SIMBAD
- ~150 terms
- Link with the dictionary of nomenclature

SIMBAD object types

```
14.09.08.0:      SN            SN*     SuperNova
14.09.09.0:      Symbiotic*    Sy*     Symbiotic Star
14.14.00.0:    Sub-stellar     su*     Sub-stellar object
14.14.02.0:      Planet?       Pl?     Extra-solar Planet Candidate
15.00.00.0:  Galaxy            G       Galaxy
15.01.00.0:    PartofG         PoG     Part of a Galaxy
15.02.00.0:    GinCl           GiC     Galaxy in Cluster of Galaxies
15.03.00.0:    GinGroup        GiG     Galaxy in Group of Galaxies
15.04.00.0:    GinPair         GiP     Galaxy in Pair of Galaxies
15.05.00.0:    High_z_G        HzG     Galaxy with high redshift
15.06.00.0:    AbsLineSystem   ALS     Absorption Line system
15.06.01.0:      Ly-alpha_ALS  LyA     Ly alpha Absorption Line system
15.06.02.0:      DLy-alpha_ALS DLA     Damped Ly-alpha Absorption Line system
15.06.03.0:      metal_ALS     mAL     metallic Absorption Line system
15.06.05.0:      Ly-limit_ALS  LLS     Lyman limit system
15.06.08.0:      Broad_ALS     BAL     Broad Absorption Line system
15.07.00.0:    RadioG          rG      Radio Galaxy
15.08.00.0:    HII_G           H2G     HII Galaxy
15.09.00.0:    LSB_G           LSB     Low Surface Brightness Galaxy
15.10.00.0:    AGN_Candidate   AG?     Possible Active Galaxy Nucleus
15.10.07.0:      QSO_Candidate Q?      Possible Quasar
15.11.00.0:    EmG             EmG     Emission-line galaxy
```


Nomenclature of Celestial Objects (Result I) - Mozilla
File Edit View Go Bookmarks Tools Window Help
Back
Forward
Reload
Stop
http://vizier.u-strasbg.fr/cgi-bin/Dic
Search
Print
Home
Bookmarks
mozilla.org
Latest Builds
Google
iNFO
IDiC

Dictionary of Nomenclature of Celestial Objects

CDS · Simbad · VizieR · Aladin · Catalogues · Nomenclature · Biblio · Tutorial · Developer's corner

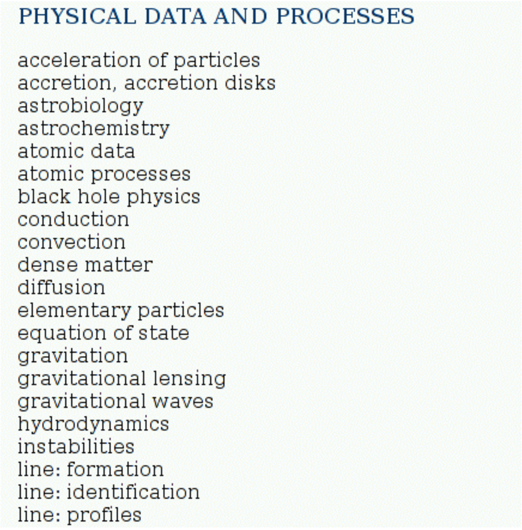
(Last CDS update: 09-Sep-2005)

Result of query: info cati **M$**

Acronym	Use	Format	Year	1st Author	Obj. Type
(LMC M)	[MYM2001]	NNN LMC M HHMMm+DDMM	2001	MIZUNO N.+	CO Cloud
M	M	NNN	1850	MESSIER C.	(Opt)
M	M	1-NN 2-NN 3-NN 4-NN	1946	MINKOWSKI R.	PN
M	M	NNN	1975	MAFFEI P.	V*
(M)	GCM	+LL.ll+BB.bb	1981	GUSTEN R.+	MCld
(M)	[GVC73]	{M} R.N	1973	GIOVANELLI R.+	Concentration
(M)	[H68]	{M} {M} R	1968	HULSBOSCH A.N.M.	HVC
(M)	[M59]	NN	1959	MANOVA G.A.	Em*
(M)	[M61a]	NN	1961	MINKOWSKI R.	G in ClG
(M)	[MAG95]	NNN	1995	MINNITI D.+	GCl
(M)	Mills	HH+{D}A	1952	MILLS B.Y.	(Rad)
(M)	[MLV92]	NNNNNN	1992	MAGNIER E.A.+	*
(M)	MM	NN	1965	MORAN M.	(Rad)
(M)	MSH	HH+D-NN	1958	MILLS B.Y.+	(Rad)

Keywords from journals

- Normalized list of keywords for the main astronomical journals
- Classification of bibliographical references



PHYSICAL DATA AND PROCESSES

acceleration of particles
accretion, accretion disks
astrobiology
astrochemistry
atomic data
atomic processes
black hole physics
conduction
convection
dense matter
diffusion
elementary particles
equation of state
gravitation
gravitational lensing
gravitational waves
hydrodynamics
instabilities
line: formation
line: identification
line: profiles

Challenges

- Combine these knowledge bases to perform complex queries
- Combine several sources of (meta)data : catalogues, journals, ...
- Help astronomers to find resources in a growing list of heterogeneous sources
  - Virtual Observatory's Registry
  - Advanced information retrieval, metadata mining

Towards Ontologies

- What technologies ?
  - Semantic Web
  - Ontologies
- How they connect to the VO
  - Registry, UCD, ADQL
- What components can be built?
- Suggested implementations

What technologies?

- Ontologies and Semantic Web are very active research domains
- XML-based
- Strong support by W3C – definition of open standards (approved after years of discussions!)
- Interest of many commercial companies in these techniques

Ontologies

- An ontology is a formal description of a set of concepts and their relationships to each other
- Why develop one ? (Ontology 101)
  - to share common understanding of the structure of information among people or software agents
  - to enable reuse of domain knowledge
  - to make domain assumptions explicit
  - to analyze domain knowledge

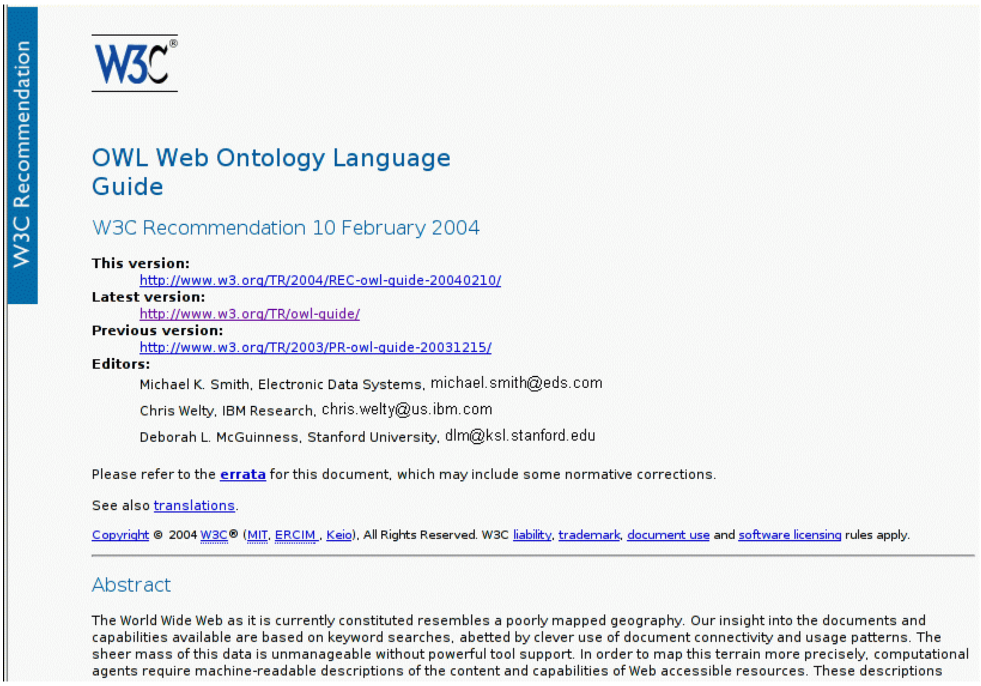
Ontologies

- Ontologies rely on Description Logics
- Reasoners can make inferences
  - defined concepts vs primitive concepts
  - necessary (sufficient) conditions
- Building an ontology is an iterative, collaborative process:
  - domain ontologies
  - task ontologies

Ontologies

- Ontologies can be edited/stored with different tools/formats:
  - DAML + OIL (Ontology Inference Layer)
  - OWL (Web Ontology Language)
- Quite mature editors:
  - OILed
  - Protégé

Ontologies



W3C Recommendation

OWL Web Ontology Language Guide

W3C Recommendation 10 February 2004

This version:
http://www.w3.org/TR/2004/REC-owl-guide-20040210/
Latest version:
http://www.w3.org/TR/owl-guide/
Previous version:
http://www.w3.org/TR/2003/PR-owl-guide-20031215/
Editors:
Michael K. Smith, Electronic Data Systems, michael.smith@eds.com
Chris Welty, IBM Research, chris.welty@us.ibm.com
Deborah L. McGuinness, Stanford University, dlm@ksl.stanford.edu

Please refer to the **errata** for this document, which may include some normative corrections.

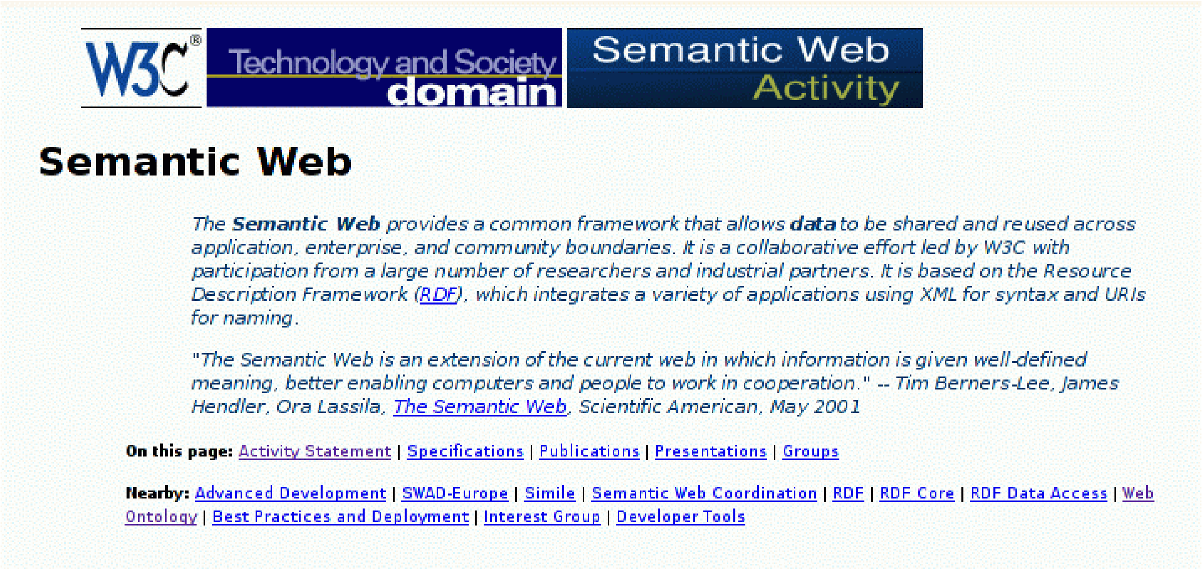
See also translations.

Copyright © 2004 W3C® (MIT, ERCIM, Keio), All Rights Reserved. W3C liability, trademark, document use and software licensing rules apply.

Abstract

The World Wide Web as it is currently constituted resembles a poorly mapped geography. Our insight into the documents and capabilities available are based on keyword searches, abetted by clever use of document connectivity and usage patterns. The sheer mass of this data is unmanageable without powerful tool support. In order to map this terrain more precisely, computational agents require machine-readable descriptions of the content and capabilities of Web accessible resources. These descriptions

Semantic Web



Semantic Web

*The **Semantic Web** provides a common framework that allows **data** to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by W3C with participation from a large number of researchers and industrial partners. It is based on the Resource Description Framework (RDF), which integrates a variety of applications using XML for syntax and URIs for naming.*

"The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." -- Tim Berners-Lee, James Hendler, Ora Lassila, The Semantic Web, Scientific American, May 2001

On this page: Activity Statement | Specifications | Publications | Presentations | Groups

Nearby: Advanced Development | SWAD-Europe | Simile | Semantic Web Coordination | RDF | RDF Core | RDF Data Access | Web Ontology | Best Practices and Deployment | Interest Group | Developer Tools

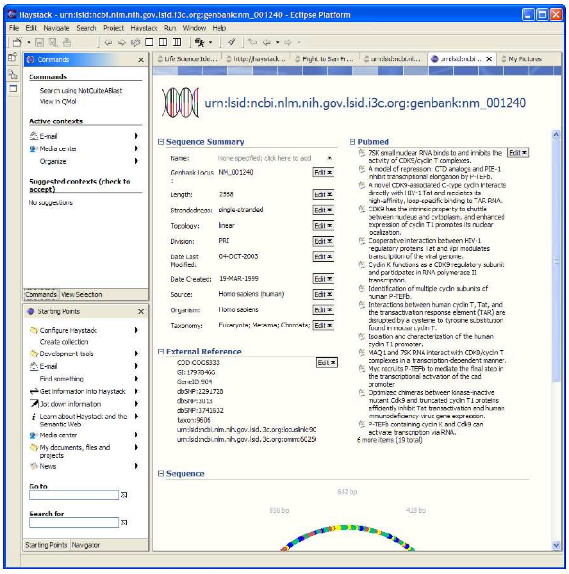
Semantic Web

- Data shared and reused among application and community boundaries
- RDF: Resource Description Framework (XML-based, with URIs and namespaces)
- Goal:
  - communication between software agents, users
  - discover valuable applications (resource discovery) and exploitation paths (workflows)

Application in biology

- BioHaystack (IBM Watson Research)

- Integration of data from heterogeneous databases
  - access protocols
  - data formats
  - softwares
- RDF as underlying model
- Link to myGRID



urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:genbank:nm_001240

Sequence Summary

Name: None specified; click here to add
Genbank Locus: NM_001240
Length: 2558
Strandedness: single-stranded
Topology: linear
Division: PRI
Date Last Modified: 04-OCT-2003
Date Created: 19-MAR-1999
Source: Homo sapiens (human)
Organism: Homo sapiens
Taxonomy: Eukaryota; Metazoa; Chordata;

External Reference

CDD:COG5333
GI:17978466
GeneID:904
dbSNP:2291728
dbSNP:3013
dbSNP:3741632
taxon:9606
urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:locuslink:90
urn:lsid:ncbi.nlm.nih.gov.lsid.i3c.org:omim:6025

Pubmed

- 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes.
- A model of repression: CTD analogs and PIE-1 inhibit transcriptional elongation by P-TEFb.
- A novel CDK9-associated C-type cyclin interacts directly with HIV-1 Tat and mediates its high-affinity, loop-specific binding to TAR RNA.
- CDK9 has the intrinsic property to shuttle between nucleus and cytoplasm, and enhanced expression of cyclin T1 promotes its nuclear localization.
- Cooperative interaction between HIV-1 regulatory proteins Tat and Vpr modulates transcription of the viral genome.
- Cyclin K functions as a CDK9 regulatory subunit and participates in RNA polymerase II transcription.
- Identification of multiple cyclin subunits of human P-TEFb.
- Interactions between human cyclin T, Tat, and the transactivation response element (TAR) are disrupted by a cysteine to tyrosine substitution found in mouse cyclin T.
- Isolation and characterization of the human cyclin T1 promoter.
- MAQ1 and 7SK RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner.
- Myc recruits P-TEFb to mediate the final step in the transcriptional activation of the cad promoter
- Optimized chimeras between kinase-inactive mutant Cdk9 and truncated cyclin T1 proteins efficiently inhibit Tat transactivation and human immunodeficiency virus gene expression.
- P-TEFb containing cyclin K and Cdk9 can activate transcription via RNA.

6 more items (19 total)

Sequence

Application to the VO

- Astronomy is a user community which is familiar with many issues motivating the Semantic Web research:

- Resource description in the VORegistry, with unique identifiers and XML format: mapping to RDF?
- VOTable, with metadata and data grouped in the same document: metadata sharing is at the core of the semantic web
- UCD and the IAU thesaurus for the semantic description will be very valuable resources for building an astronomical ontology

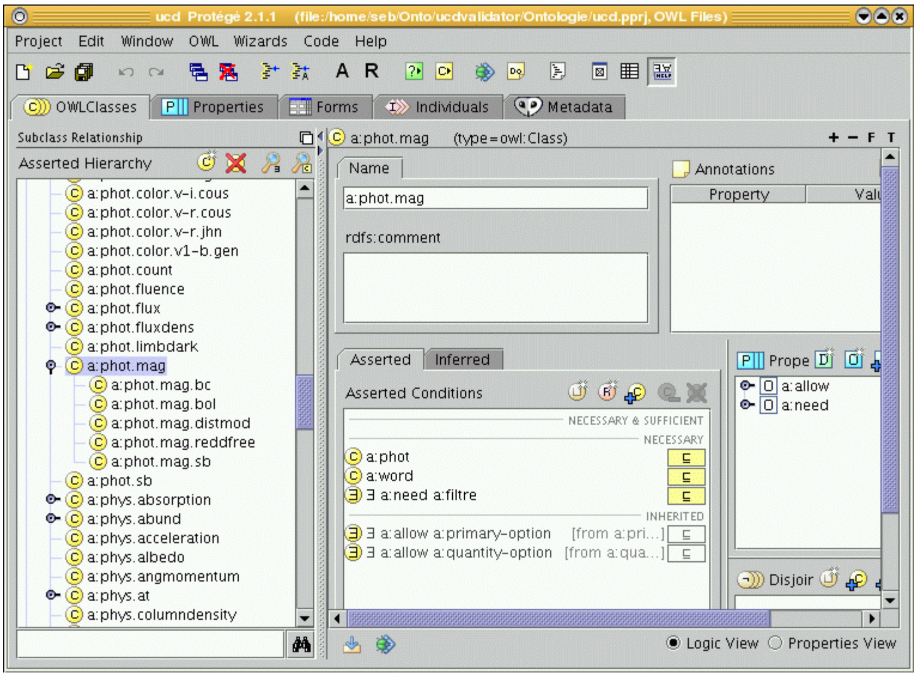
How to proceed ?

- Don't only try to build one single ontology to describe all astronomy
- Start with well-defined small ontologies, with simple use cases
- Then, build links between these ontologies and allow more complex problems to be solved

UCD and ontologies

- Collaboration between CDS and LORIA
- Syntactic rules for building UCDs from words
- UCD validation
- Application to UCD assignment:
  - given a description, which UCD is relevant?

UCD and ontologies



ucd Protégé 2.1.1 (file:/home/seb/Onto/ucdvalidator/Ontologie/ucd.pprj, OWL Files)

a:phot.color.v-i.cous
a:phot.color.v-r.cous
a:phot.color.v-r.jhn
a:phot.color.v1-b.gen
a:phot.count
a:phot.fluence
a:phot.flux
a:phot.fluxdens
a:phot.limbdark
a:phot.mag
a:phot.mag.bc
a:phot.mag.bol
a:phot.mag.distmod
a:phot.mag.reddfree
a:phot.mag.sb
a:phot.sb
a:phys.absorption
a:phys.abund
a:phys.acceleration
a:phys.albedo
a:phys.angmomentum
a:phys.at
a:phys.columndensity

a:phot.mag (type=owl:Class)

Name: a:phot.mag

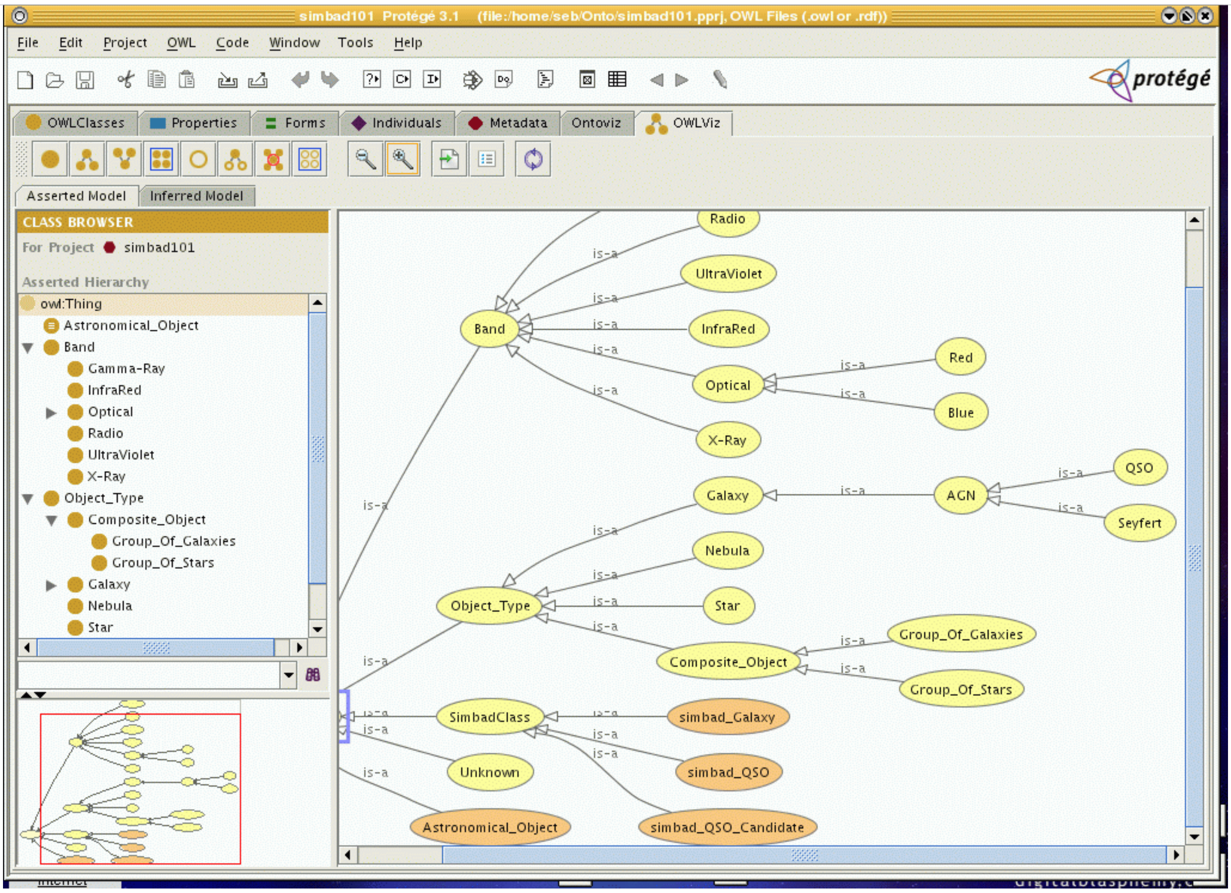
rdfs:comment

a:phot ⊑
a:word ⊑
∃ a:need a:filtre ⊑

∃ a:allow a:primary-option [from a:pri...] ⊑
∃ a:allow a:quantity-option [from a:qua...] ⊑

Ontology of object types

- Collaboration CDS-INAF in VOTech
- Based on SIMBAD object types
  - improve object hierarchical description
  - link between object type and wl/frequency domain
- Allow multiple classification (possible in SIMBAD4)
- Link with UCD and IAU thesaurus



Use case 1

- Use ontology to validate entries during registry population (for data curators):
  - interactive keywords refinement
  - check if global keywords consistent with specific measurements (UCDs)

Use case 2

- Assist users in registry exploration
- Goal: retrieve relevant datasets from the registry, using the ontology
  - search for "young star" would be interpreted by a reasoner, and recursively matched to all sub-categories of "young star"
  - in case of null result, suggest request broadening (TTauri star -> young star)

Use case 3

- Help automated object names detection in articles
  - analyze the context of the paper
    - keywords
    - object types
    - measurements (UCDs)
  - use dictionary of nomenclature
  - identify object names, and most likely acronym

Questions ?