Exploiting Virtual Observatory and Information Technology: Techniques for Astronomy

Lecture #2 Goal:

Data Centres and Databases

Discovery, access, federating

Nicholas Walton
AstroGrid Project Scientist
Institute of Astronomy,
The University of Cambridge

Summary: Lecture #2

- Introduction
- Science Archives
  - MetaCentres
  - Missions
- Databases
  - XML and Registries
  - Queries, SQL
- Federating Databases
  - Cross matching
  - Open Sky Query/ Open Sky Nodes/ Data Set Access
- Science Example
  - Hunting for Brown Dwarfs

Introduction: Catalogue Access

- Data exists in many globally located archives
  - In addition to data on your tapes
- How to find and access that data
  - Issues of types of data
  - Issues of types of databases and access to data
  - Issues of data description
- Virtual Observatory standards to address these
  - VOTable, UCD, Registry, VOQL
- Technologies
  - XML, SQL
- Note: many of the concepts and standards referred to here are rapidly evolving, so information in this lecture may soon be out of date!

Missions & Data Centres: Overview

Creating the digital sky...

Traditional Data Centres: USA

- Radio:
  - NRAO: http://e2e.nrao.edu/archive/
- InfraRed:
  - IPAC @ CalTech: http://www.ipac.caltech.edu/
- Optical:
  - MAST @ STScI: http://archive.stsci.edu/mast.html
  - Observatories: SDSS http://www.sdss.org, NOAO http://www.archive.noao.edu/nsa/, Keck http://www2.keck.hawaii.edu/koa/koa.php
- UV/X-Ray:
  - HEASARC @ GSFC: http://heasarc.gsfc.nasa.gov/
  - Chandra @ SAO: http://cxc.harvard.edu/cda/
- γ-Ray: Swift @ HEASARC
- Solar: see NSO @ http://vso.nascom.nasa.gov/cgi-bin/search

Data Centres: UK

- Radio:
  - Jodrell Bank/Merlin: http://www.merlin.ac.uk/archive/
- InfraRed/ Optical:
  - CASU @ IoA, Cambridge: http://archive.ast.cam.ac.uk/
  - WFAU @ ROE, Edinburgh: http://www.roe.ac.uk/ifa/wfau/
  - UK ISO: http://jackal.bnsc.rl.ac.uk/isouk/ (ends 2006)
- Xray & γ-Ray:
  - LEDAS @ Leicester: http://ledas-www.star.le.ac.uk/
- Solar:
  - RAL: http://trace.solararchive.rl.ac.uk/soho/
  - MSSL: http://www.mssl.ucl.ac.uk/www_solar/surfindex.html
- STP:
  - RAL WDC: http://www.wdc.rl.ac.uk/
  - Lancaster: http://www.dcs.lancs.ac.uk/iono/data/

Data Centres: EU

- Radio: see also Radionet @ http://www.radionet-eu.org/
  - JIVE: http://archive.jive.nl/scripts/listarch.php
  - ESO: will host ALMA sub-mm archive
- IR/ Optical: see also Opticon @ http://www.astro-opticon.org/
  - ESO (ground): http://archive.eso.org/
  - OmegaCen @ Groningen: http://www.astro.rug.nl/~omegacen/
  - Terapix (CFHT/MegaCam): http://terapix.iap.fr/rubrique.php?id_rubrique=169
- ESA (space missions):
  - http://www.rssd.esa.int/index.php?project=SA
- High Energy
  - XMM: http://xmm.vilspa.esa.es/external/xmm_data_acc/xsa/
- Solar/ STP: no one major collection (except see ESA)

Data Centres: Other

- ADC @ NAOJ http://dbc.nao.ac.jp/
  - Subaru @ SMOKA http://smoka.nao.ac.jp/
- CFHT @ CADC
  - http://cadcwww.dao.nrc.ca/cfht/
- Gemini @ CADC
  - http://cadcwww.hia.nrc.ca/gemini/

Meta Centres

- NED: http://nedwww.ipac.caltech.edu/
- CDS: http://cdsweb.u-strasbg.fr/
  - Vizier: Catalogues
  - Simbad:
  - Aladin: Integration and Visualisation
- Planetary Data Centre: http://pds.jpl.nasa.gov/
- CADC: http://cadcwww.dao.nrc.ca/
- Astro-ph: http://uk.arxiv.org/archive/astro-ph
  - Pre-print server
- ADS: http://ukads.nottingham.ac.uk/
  - Publications
- Google: http://www.google.com
  - Google Scholar http://www.scholar.google.com/

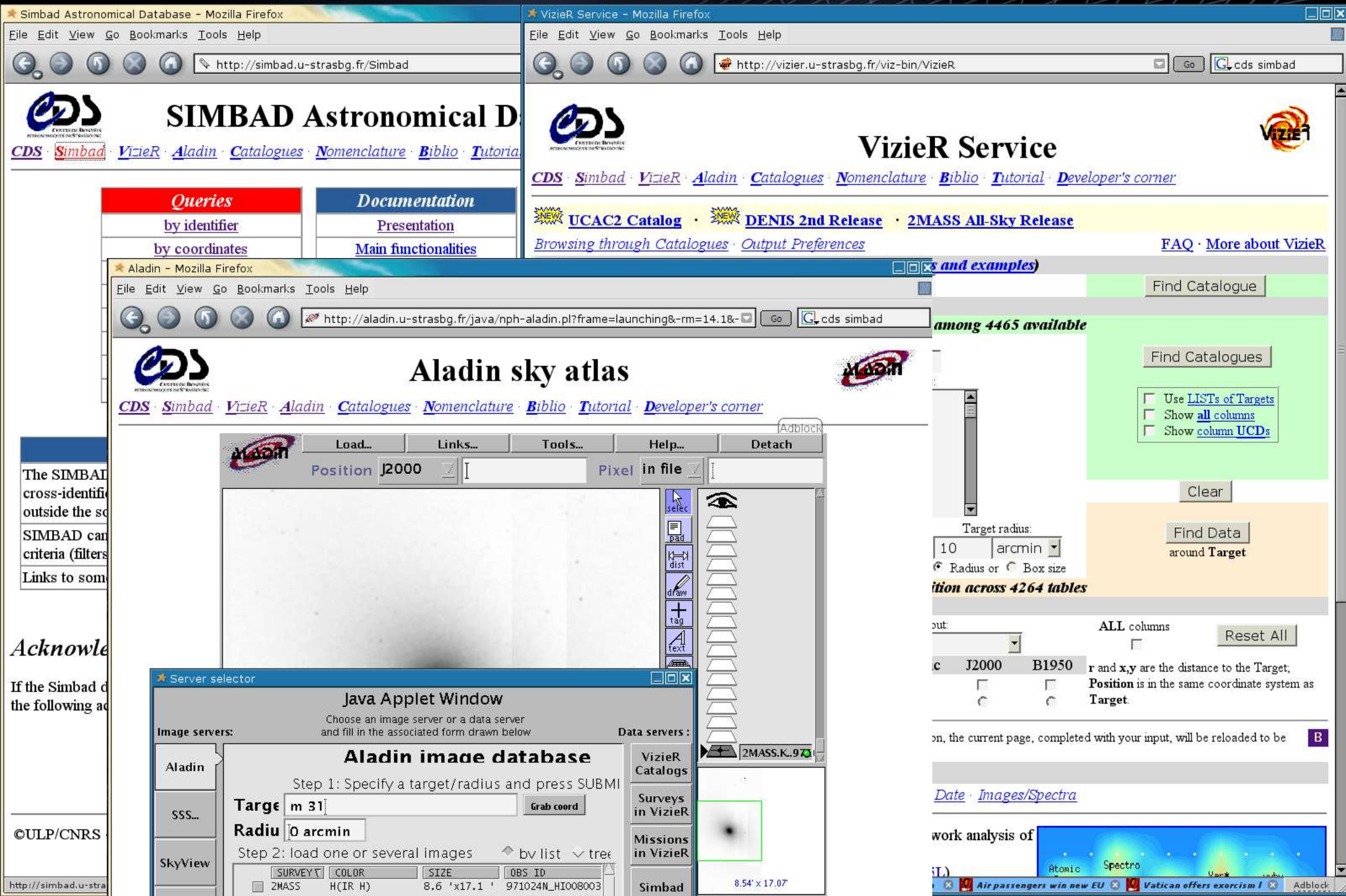
Athens, the Library, Computing Course

- Eduserv Athens: Useful resource
  - Login @ http://www.athensams.net/myathens/
  - Entry way to Blackwell, Ingenta, ISI Web of Science
- Cambridge Library
  - IoA: http://www.ast.cam.ac.uk/~ioalib/homepage.html
  - Database links: http://www.ast.cam.ac.uk/~ioalib/databases.html
  - Newton: http://newton.lib.cam.ac.uk:7603/
- IoA Graduate Computing Course
  - Recap: Jeremy Saunders computing course
    - Notes @ http://www-xray.ast.cam.ac.uk/~jss/lecture/grad_training/notes/
    - Resources @ http://www-xray.ast.cam.ac.uk/~jss/lecture/grad_training/

NED: A resource for 8 Million **ExtraGalactic** Objects

http://nedwww.ipac.caltech.edu/

CDS: http://cdsweb.u-strasbg.fr/



How to find and query data with VOs

- Problem: there is obviously a lot of data available at individual archive sites
- But, how does one find relevant data for say an individual object, a patch of sky, a set of galaxies with a certain morphological class?
- How does one avoid having to access each of these individual archives, one at a time.
- Solution: the Virtual Observatory, and its underpinning interoperability standards
  - A 'one stop' solution ...
- But first, XML, Registries, and SQL ...

XML: Structured Information

- Extensible Markup Language for Documents and Data
- Readable and verbose
- Schema which set structure, used for data models
- Transformable (using XLST)
  - XML, HTML, PDF and so forth
- Tools to create and debug readily available
  - Parsers: Java, C++, Perl ...
  - Browsers and Editors
  - Databases in XML
    - Exist (used by AstroGrid): http://exist.sourceforge.net/
  - Bindings > APIs

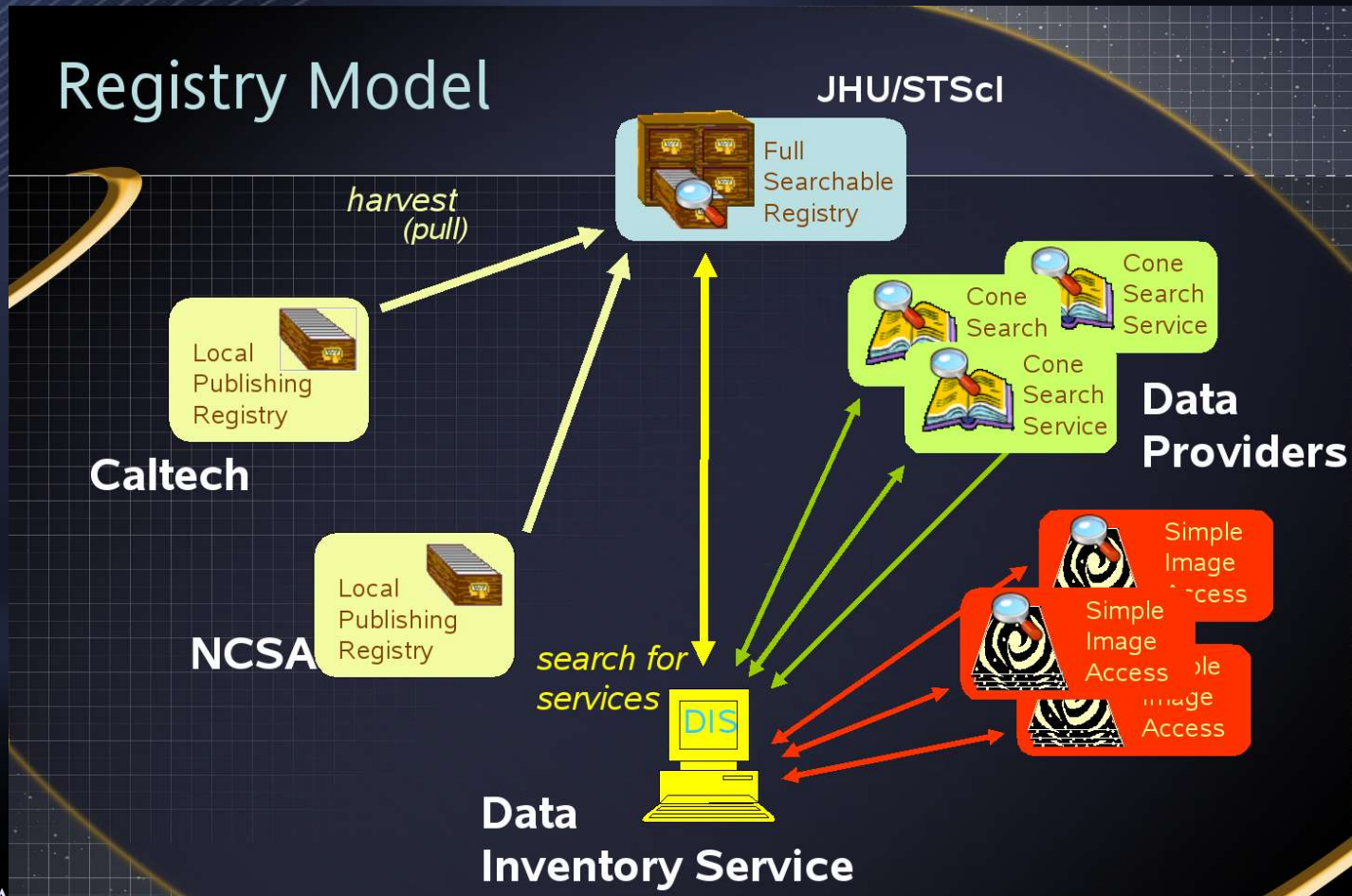
http://xml.oreilly.com

Locating Relevant Data

- As an astronomer – how do you find the data that you require?
- Are their multiple resources available?
- How do you decide which of the resources is actually relevant?
- Resources can be:
  - Data
  - Information (e.g. Paper references)
  - Applications or other programmes
  - Compute/ disk etc
- VO registries provide a solution to these questions

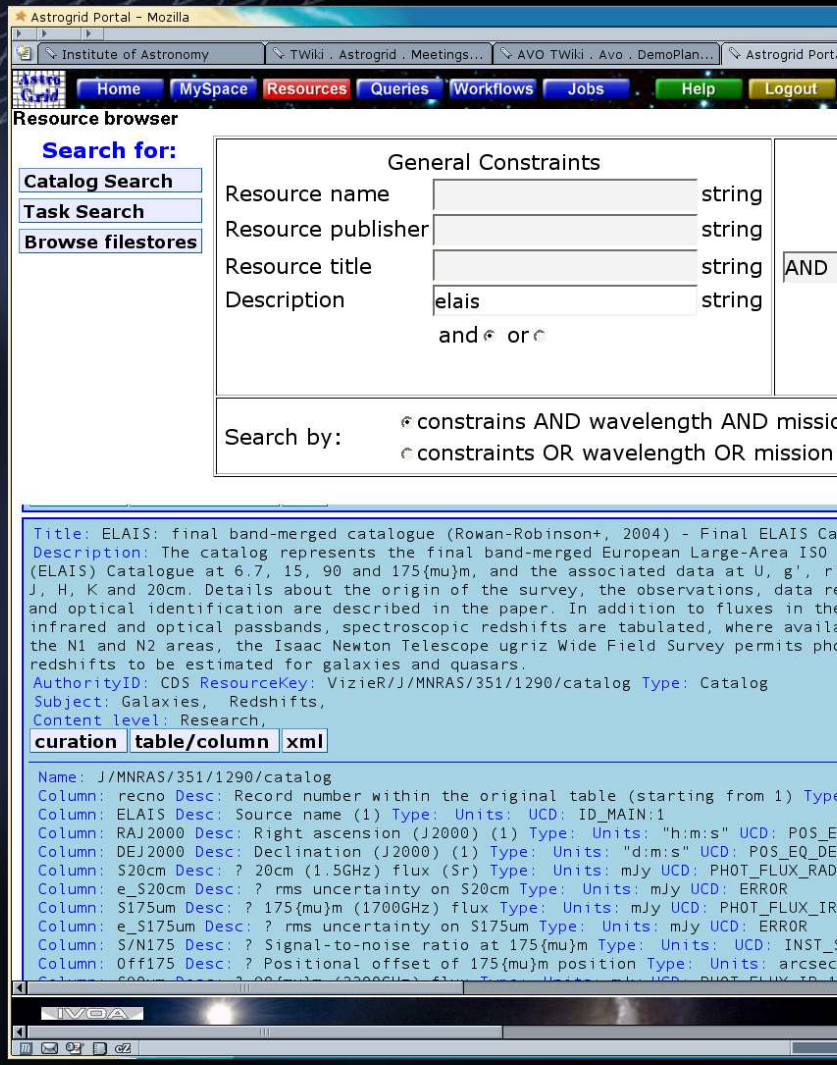
Registries: what are they?

- Used to discover and locate resources
- A list of resource descriptions, described by structured metadata: enables automated searching and processing
- Resource metadata
  - XML schemas



AstroGrid's Registry

- Types of Registries:– Full, Publish, Special
- Registry is the main focal point for all Astrogrid components
- Agreed Standards with IVOA
  - Search and harvest interface
  - OAI (a digital library) standard for harvest interface
- Types of resource
  - Generic services, web services, applications, …
  - Data collections
  - AstroGrid-specific resources (e.g. MySpace servers)
- Use of XQuery language with eXist XML database.
- Harvesting (Jan 05)
  - US NVO, CDS-VizieR



VO registries

- NVO registry
- AstroGrid registry
- Euro-VO registry (based on AstroGrid implementation)
- CDS registry
- Japan-VO registry

All now harvesting each other: thus querying any one returns full list of globally held resources.

VOResource XML Schema:

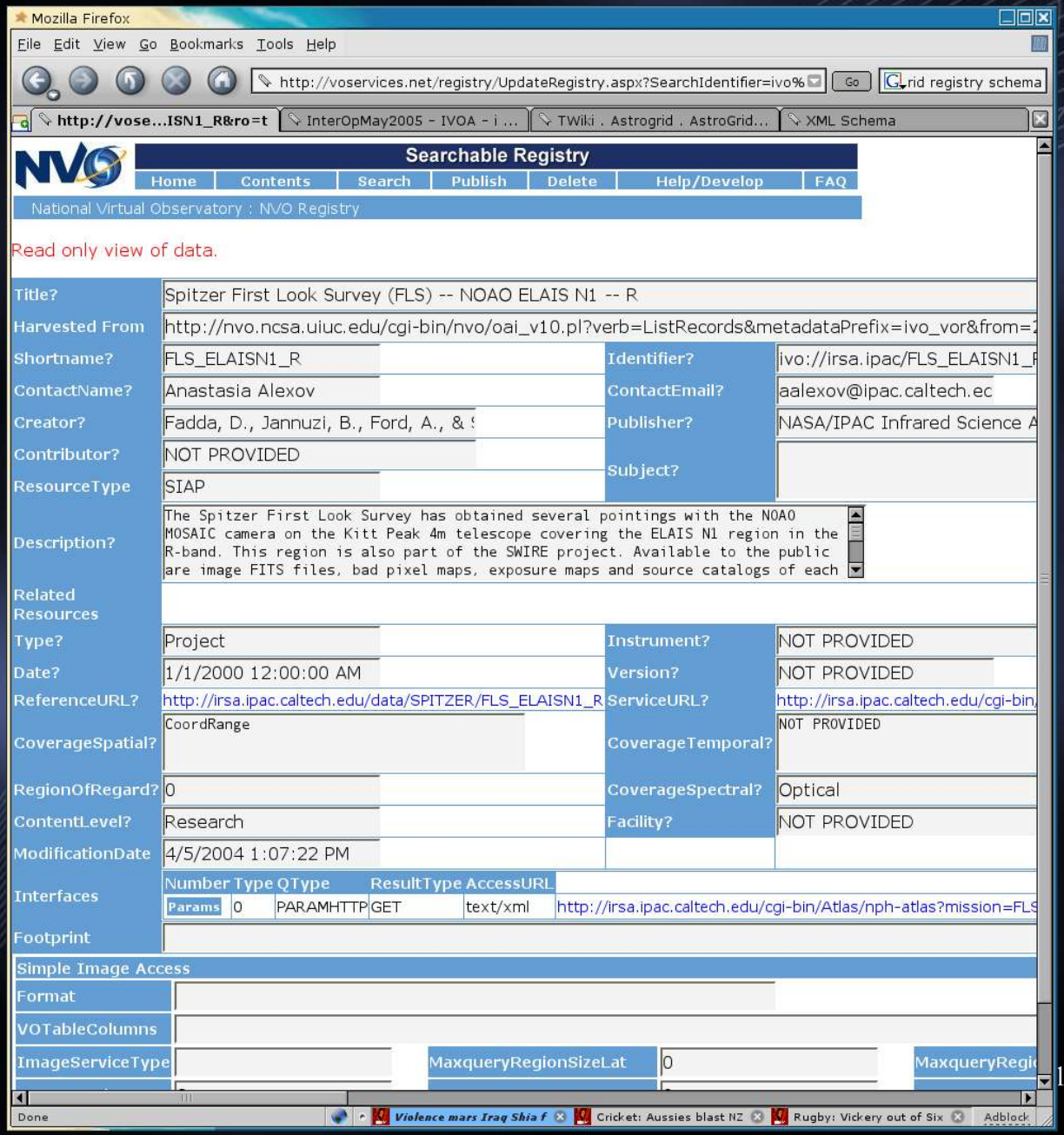
http://www.ivoa.net/xml/VOResource/v0.10

```
<?xml version="1.0" encoding="UTF-8"?>
<xs:schema targetNamespace="http://www.ivoa.net/xml/VOResource/v0.10"
           xmlns="http://www.w3.org/2001/XMLSchema"
           xmlns:xs="http://www.w3.org/2001/XMLSchema"
           xmlns:vr="http://www.ivoa.net/xml/VOResource/v0.10"
           xmlns:vm="http://www.ivoa.net/xml/VOMetadata/v0.1"
           elementFormDefault="qualified"
           attributeFormDefault="unqualified"
           version="0.10">

<xs:include schemaLocation="VOResourceRelType-v0.10.xsd"/>

<xs:complexType name="Resource">
      <xs:annotation>
         <xs:documentation>
           Any entity that is describable and identifiable by a IVOA Identifier.
         </xs:documentation>
      </xs:annotation>
      <xs:sequence>
         <xs:element name="title" type="xs:string">
            <xs:annotation>
               <xs:appinfo>
                 <vm:dcterm>Title</vm:dcterm>
               </xs:appinfo>
               <xs:documentation>
                  the full name given to the resource
               </xs:documentation>
            </xs:annotation>
         </xs:element>
```

Web Page to Registry: Use of Schemas



- Registry Schema define structure
- This example shows a resource described using the VOResource schema
- Registry populated via a mixture of automatic and manual entry of resource information

VOTable: An interchange format

Source: VOTable 1.1: http://www.ivoa.net/Documents/REC/VOTable/VOTable-20040811.html

The VOTable format is an XML standard for the interchange of data represented as a set of tables. In this context, a table is an unordered set of rows, each of a uniform format, as specified in the table *metadata*. Each row in a table is a sequence of table cells, and each of these contains either a primitive data type, or an array of such primitives. VOTable is derived from the Astrores format [1], itself modeled on the FITS Table format [2]; VOTable was designed to be closer to the FITS Binary Table format.

VOTable supports separating of data from metadata and the streaming of tables, and other ideas from modern distributed computing. It bridges two ways to express structured data: XML and FITS. It tries (through the UCD - see below) to express formally the semantic content of a parameter or field. It has the hierarchy and flexibility of XML: using GROUP elements introduced in version 1.1, columns in a VOTable can be grouped in arbitrarily complex hierarchies; and the ID attribute can be used in XML to enable what are essentially pointers. FITS does not handle Unicode (extended alphabet) characters.

- Full metadata representation
- A hierarchy of RESOURCEs containing PARAMs and TABLEs
- Use of UCDs o express the content of a parameter
- Metadata first (XML) then the data (XML, binary, FITS)
  - Streaming allowed with binary data

VOTable: An Example

```
<?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" "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>
```

UCDs: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaUCD

- Unified Content Descriptors: Controlled vocabulary for Astronomy

In the semantic web community, knowledge is represented through triples of well-defined words, representing *a property of an instance of a concept and its value*. For example in the statement: "The velocity of the M31 galaxy is 120 km/sec", we have a property (velocity), a concept (galaxy), an instance of the concept (the M31 galaxy), and the value (120 km/sec).

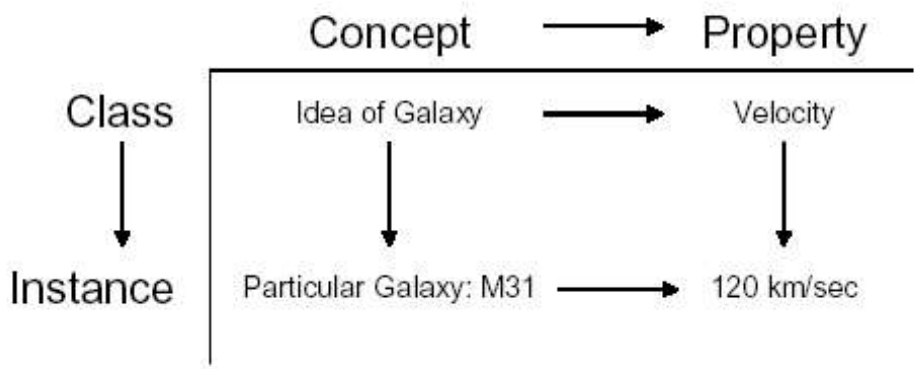


Figure 1: How concepts, properties, classes and instances are related

The UCD system is an attempt to describe simply the most commonly used quantities that astronomers want to exchange. It gives ***standard names to properties of instances of concepts*** (read this sentence twice). In Figure 1, we have:

The primary word of the UCD refers to a property, e.g.

- `meta.id`
- `meta.note`
- `phot.flux`
- `phys.mass`

Secondary words can indicate either:

- A concept that the property refers to:
  - `meta.id;src`
  - `meta.id;inst.telescope`
- Another property that the primary word refers to:
  - `stat.error;pos.eq.ra`
- Or information related to the primary word:
  - `phot.flux;em.opt.V.Johnson`

SQL

Sams Teach Yourself Sql in 10 Minutes

- Structured Query Language: an ANSI standard language designed for manipulation of relational databases. Initially developed by Codd at IBM, early 1970's.
- Latest ANSI standard is SQL2003
  - Various flavours of SQL: IBM DB2, MySQL, Oracle Database 10g, PostgreSQL, Microsoft SQL Server 2000
  - All 'more or less' implement SQL2003, but with varying syntaxes and additional Statements and (especially) Functions

```
/* MySQL */
SELECT CURRENT_TIMESTAMP;
'2001-12-15 23:50:26'
```

```
/* DB2 */
VALUES CURRENT_TIMESTAMP
'2001-12-15 23.50.26.000000'
```

- Simple syntax: `SELECT <cols> FROM <table> WHERE <conditions>`
- Joins for multiple tables: `SELECT g.*, n.type FROM galaxy g, name n WHERE g.id=n.id AND g.i>20.3`

ADQL: http://www.ivoa.net/twiki/bin/view/IVOA/IvoaVOQL

- Used in querying of single databases
- SQL92 with additional extensions specific for astronomy
  - Mathematical Functions
  - `REGION` keyword:
    - e.g. `REGION('Circle J2000 195.1 -0.34 2.3')`
  - `XMATCH` keyword:
    - e.g. `XMATCH(o,t,3.5)` (`o` and `t` are alias's)
    - `XMATCH` will be more fully covered in Lecture 4
- ADQL has two formats: /s (string) for us, and /x (xml)
  - Services to translate between formats (e.g. AstroGrid and NVO)
    - ADQL-0.7.4 services at http://openskyquery.net/adqltranslator/
- Latest version: ADQL-0.9 (2004-11-03)
  - http://www.ivoa.net/internal/IVOA/IvoaVOQL/WD_ADQL-0.9.pdf
  - Caution: most services still using ADQL-0.7.4!

ADQL /s and /x

```
select * FROM twomass_psc AS T1 WHERE CIRCLE('J2000', 12.34, -1.23, 0.01)
```

```
<Select xmlns='http://www.ivoa.net/xml/ADQL/v0.7.4'
xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
xmlns:xsd='http://www.w3.org/2001/XMLSchema' >
   <SelectionList>
      <Item xsi:type='allSelectionItemType' ></Item>
   </SelectionList>
   <From>
      <Table xsi:type='tableType' Name='twomass_psc' Alias='twomass_psc' ></Table>
   </From>
   <Where>
      <Condition xsi:type='regionSearchType' >
         <Region xmlns:q1='urn:nvo-region' xsi:type='q1:circleType' coord_system_id='' >
            <q1:Center ID='' coord_system_id='' >
               <Pos2Vector xmlns='urn:nvo-coords' >
                  <Name>Ra Dec</Name>
                  <CoordValue>
                     <Value>
                        <double>12.34</double>
                        <double>-1.23</double>
                     </Value>
                  </CoordValue>
               </Pos2Vector>
            </q1:Center>
            <q1:Radius>0.01</q1:Radius>
         </Region>
      </Condition>
   </Where>
</Select>
```

Note Schema: this is ADQL 0.7.4

Same query: the /s version is shorter and easier for us to read. But the XML /x version is better for computers!

SkyQuery: http://openskyquery.net

- Addresses the issue of sending a query to MULTIPLE databases
  - NVO implementation to date
  - AstroGrid implementation: UK databases accessible shortly
- Process
  - Webservice
    - Takes ADQL, returns VOTable
  - Analyse the query
  - Generate 'cost' estimates
  - Derive an execution plan
  - Perform x-matches
    - From small to large
    - propagate required attributes
  - VOTable returns

```
SELECT o.objId, o.r, o.type,
       t.objId, t.j_m
  FROM SDSS:PhotoPrimary o,
       TWOMASS:PhotoObj t
 WHERE XMatch(o,t)<3.5
       AND Region('CIRCLE J2000…')
       AND (o.i-t.j_m)>2
       AND o.type=3
```

VOQL

- Extension of SQL/ ADQL to allow richer high level queries of a wider variety of data, so images and not just catalogues
- SIAP to be covered in lecture 3

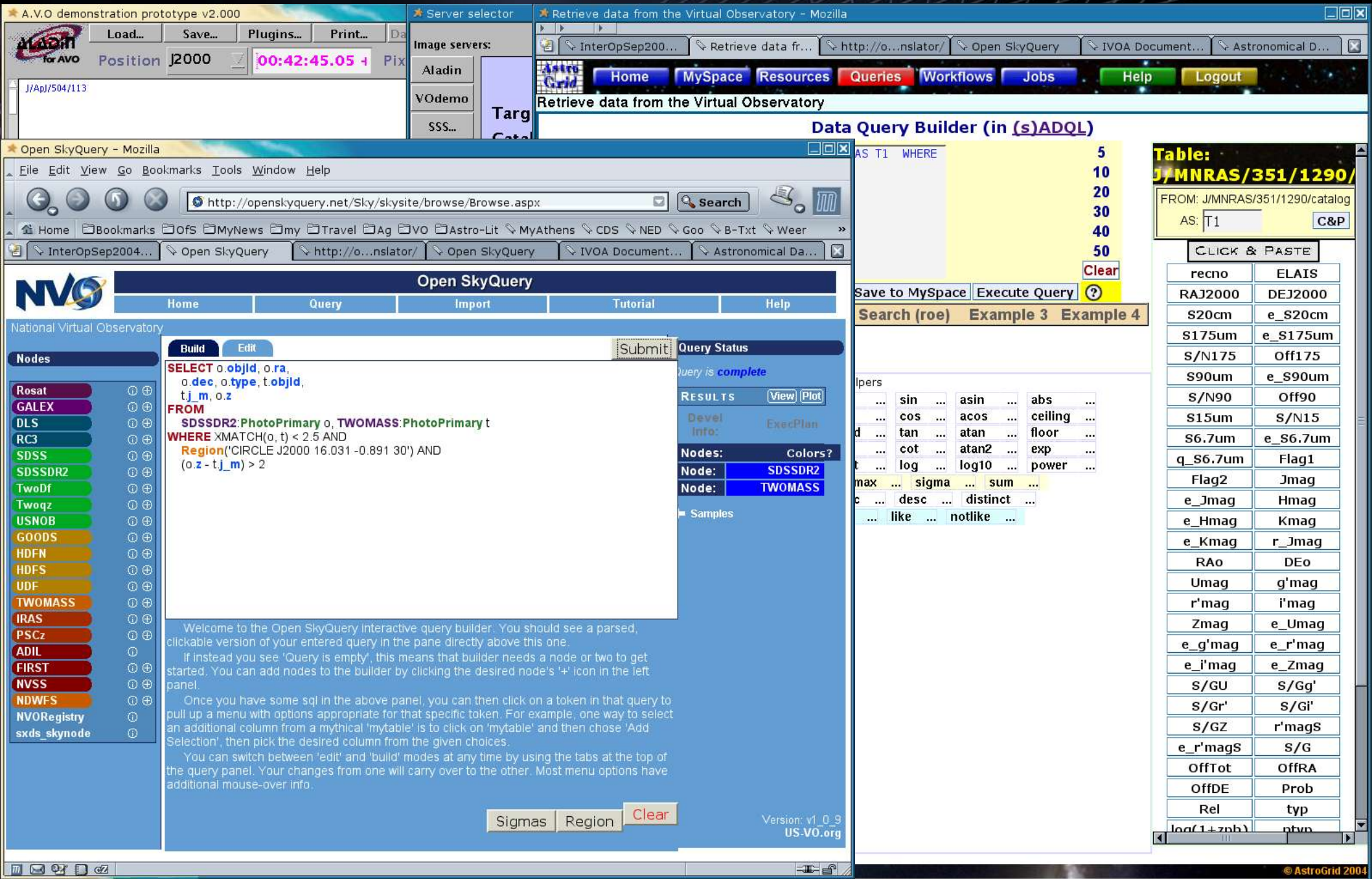
Integration of SkyNode and SIAP node

We need to make standards not only for query language but also supported protocol, the way of metadata query and specification of the returned VOTable.

	SIAP Node	SkyNode	UnifiedSkyNode
Data	Image	Catalog	Catalog/Image/Spectrum/…
Query Language	Parameter, name of the parameters are defined.	SQL, no restriction to the name of column	SQL, define a guideline for , syntax
Protocol	Http/Get/Parameter	Http/Put/Soap	Http/Put/Soap ?
Metadata Query	"FORMAT = METADATA" returns supported parameters and columns it will return.	Tables and Columns interface are defined.	Define standard metadata tal which have metadata of table and column. Query is described with SQL.
Query Output	VOTable, Some metadata must be included.	VOTable, no specification for the content	VOTable, FIELD element should have metadata stored in the metadata table.

http://www.ivoa.net/internal/IVOA/InterOpSep2004VOQL/VOQLSyntax-yshirasa.pdf

Database Queries



Science Example

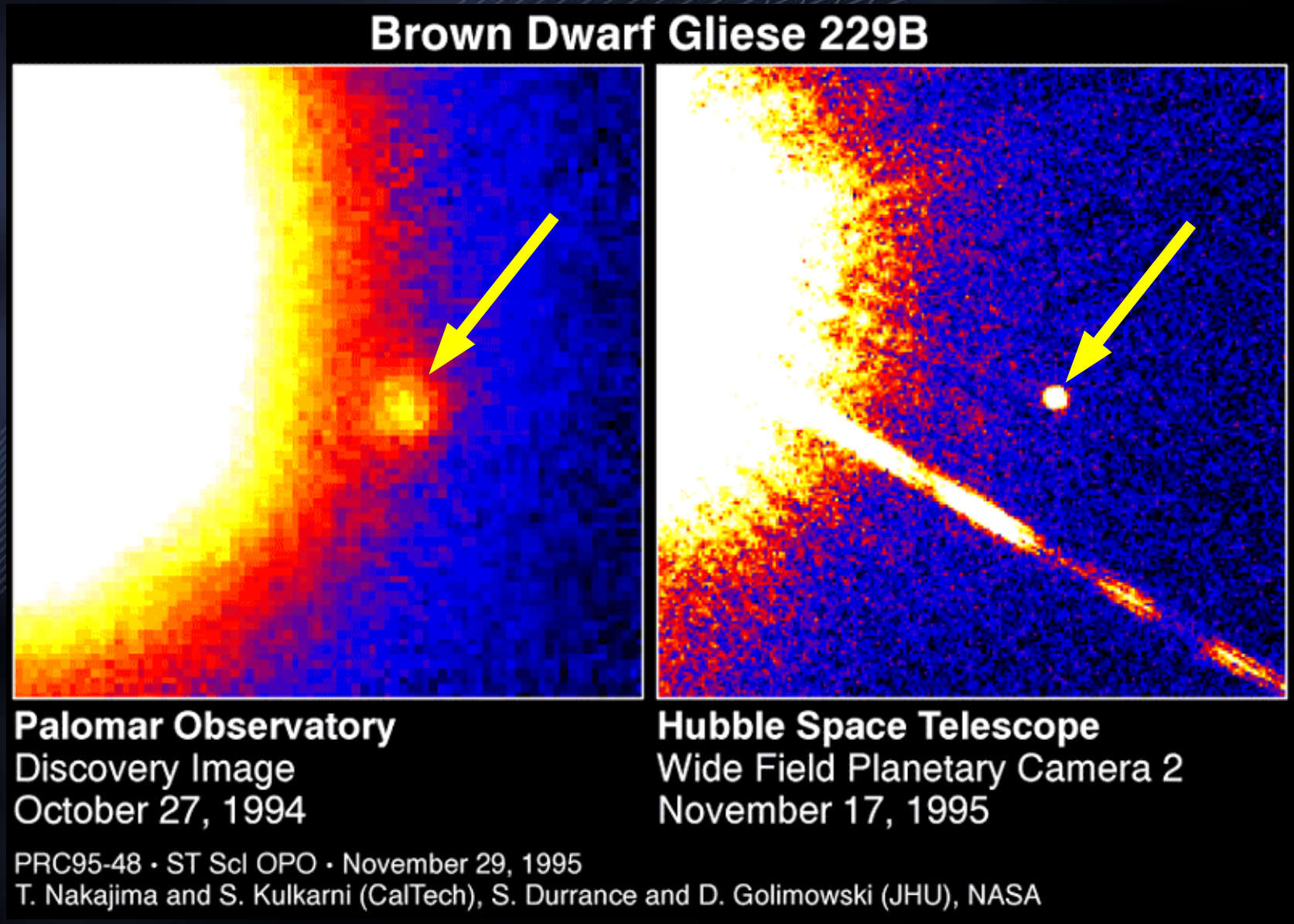
Putting the technology to use ...

Science Case: Identifying Brown Dwarfs

- Very low-mass stars
  - L-type (neutral alkali and hydride lines) and T-type (methane)
  - Cool: about 1350-2350 K
  - Inefficient nuclear fusion (mostly of deuterium)
  - Bridge the gap between stars and planets?
  - Theoretical mass range 0.012-0.08 $M_\odot$
  - *But* 'desert' - no v. large planets/small Brown Dwarfs known
    - Selection effect?
    - Different formation processes?
  - Do Brown Dwarfs form by direct cloud collapse v. planets in circumstellar discs?
  - How numerous are Brown Dwarfs ?
- Hard to find - very dim, first confirmed detection 1995

Gliese 229 B

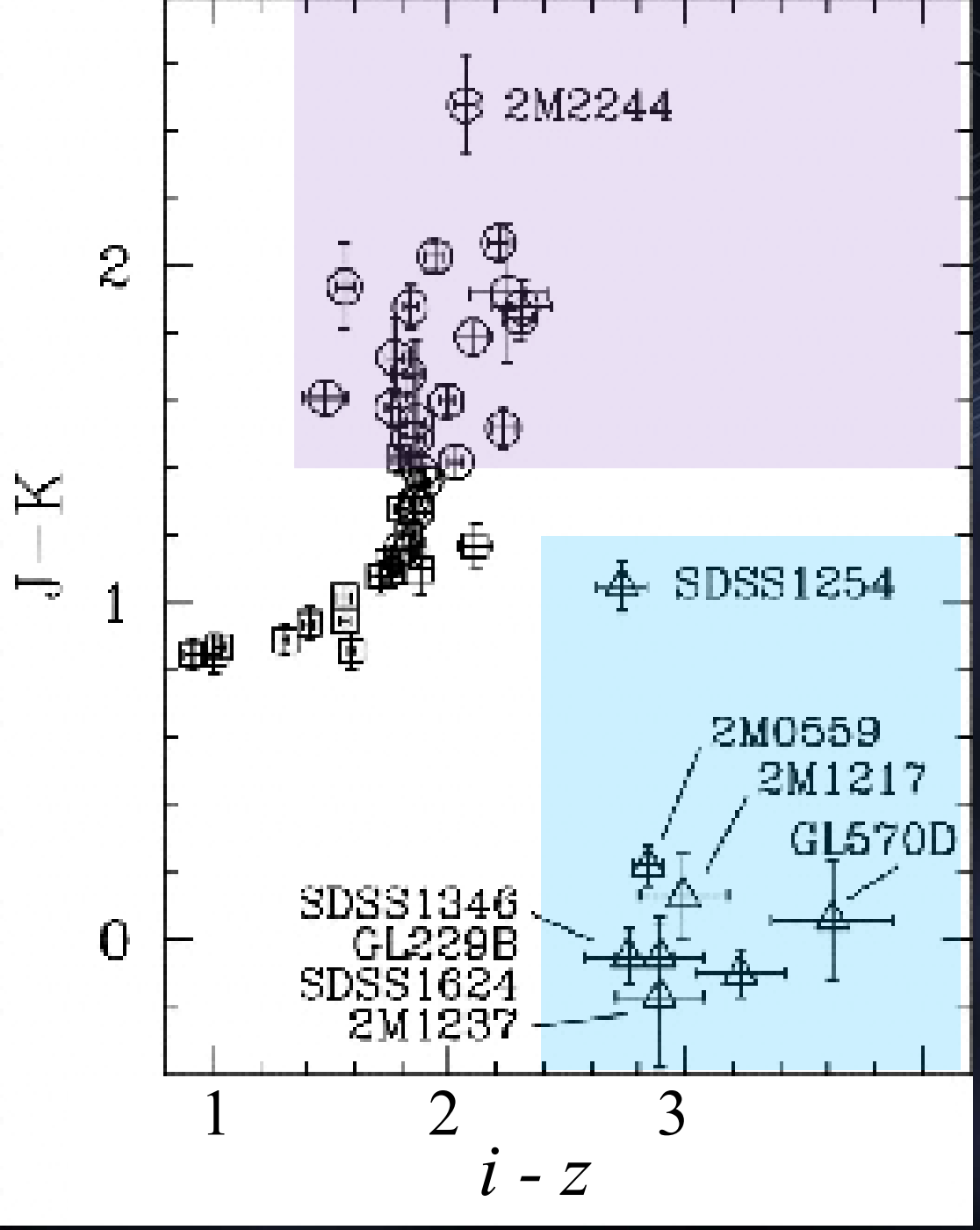
- Oldest known Brown Dwarf, most extreme colours!



Recognising very cool stars

- Most known Brown Dwarfs are identified as
  - Faint (so distance must be known)
  - Possessing distinctive spectral lines (so slow spectroscopy)
- Colours (ratios of flux density in different bands) give
  - Distance-independent temperature estimates
  - Easily obtained from large-scale surveys
- INT-WFS images give *i, z* (~770, 950 nm) bands
  - Measure *z*-band sources and *i*-band flux density at *z*-band position since upper limits are useful
- 2MASS data catalogue gives J, K (1.2, 2.2 μm) bands

Brown Dwarf colour-colour plot



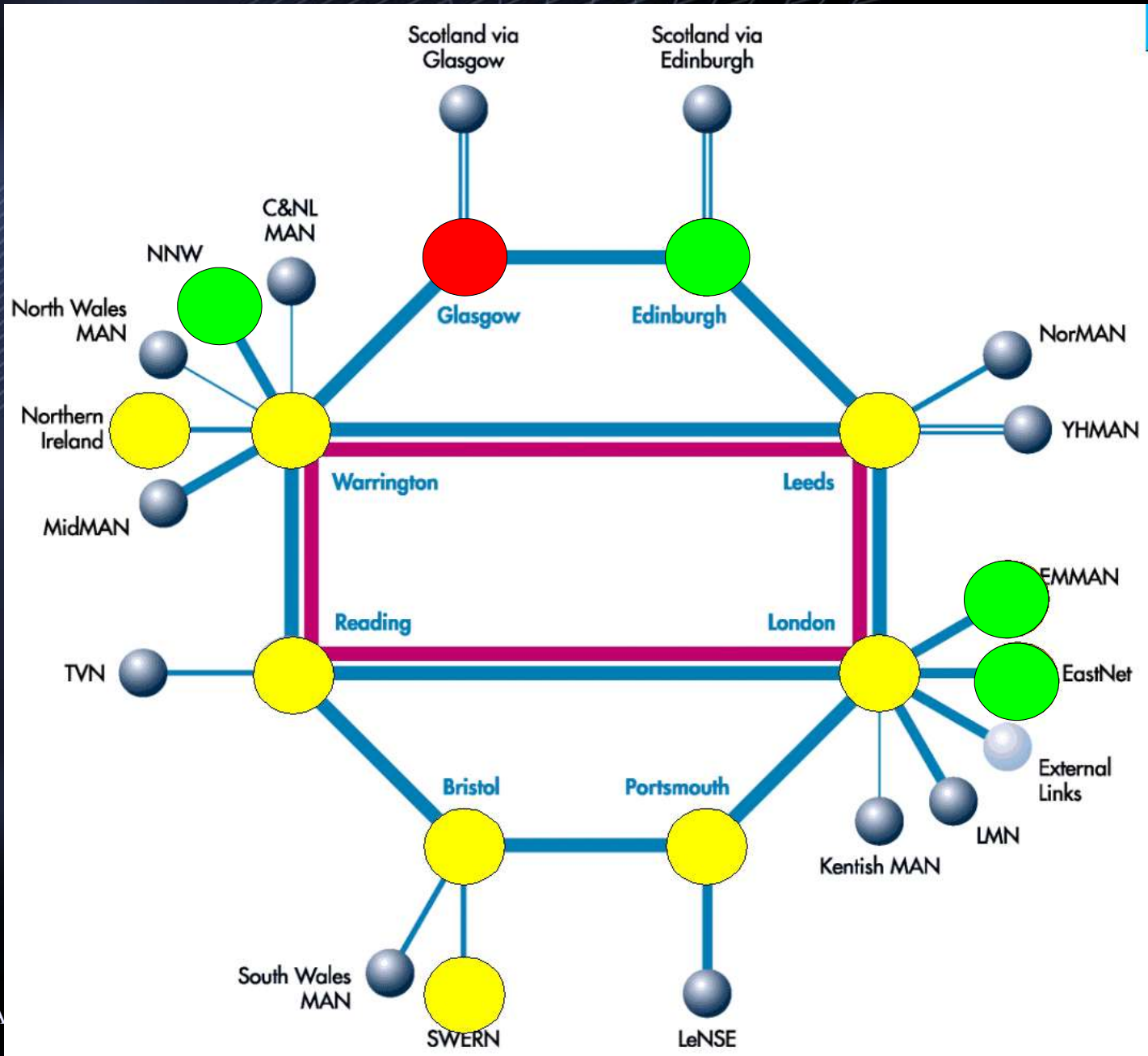
- Dahn et al. (2002)
- Dobbie et al. (2002)
  - adjust to INT-WFS filters
- ⊕ Mostly L-dwarfs
  - i-z > 1.4
  - J-K > 1.4
- △ T-dwarfs
  - i-z > 2.4
  - J-K < 1.2
- □ M-dwarfs

Accessing & processing distributed data

- Select INT-WFS Observing Log entries for Pleiades in *i* and *z* bands with small photometric and pointing errors
  - Cross-match to get *i* and *z* observations of same fields
- Extract Zero-point, Seeing, Exposure time from Logs
- Construct image URLs for Simple Image Access server at Cambridge
- Feed images to SExtractor hosted at JBO [Covered in Lecture 3]
- Cone search 2MASS catalogue at ROE for Pleiades region
- Cross-match lists of extracted i and z sources and 2MASS sources, all held in MySpace at Leicester
- Use TopCat tool to access MySpace files and make colours

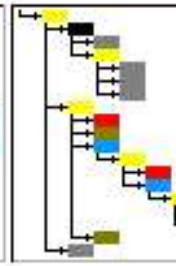
Colour cut using distributed resources

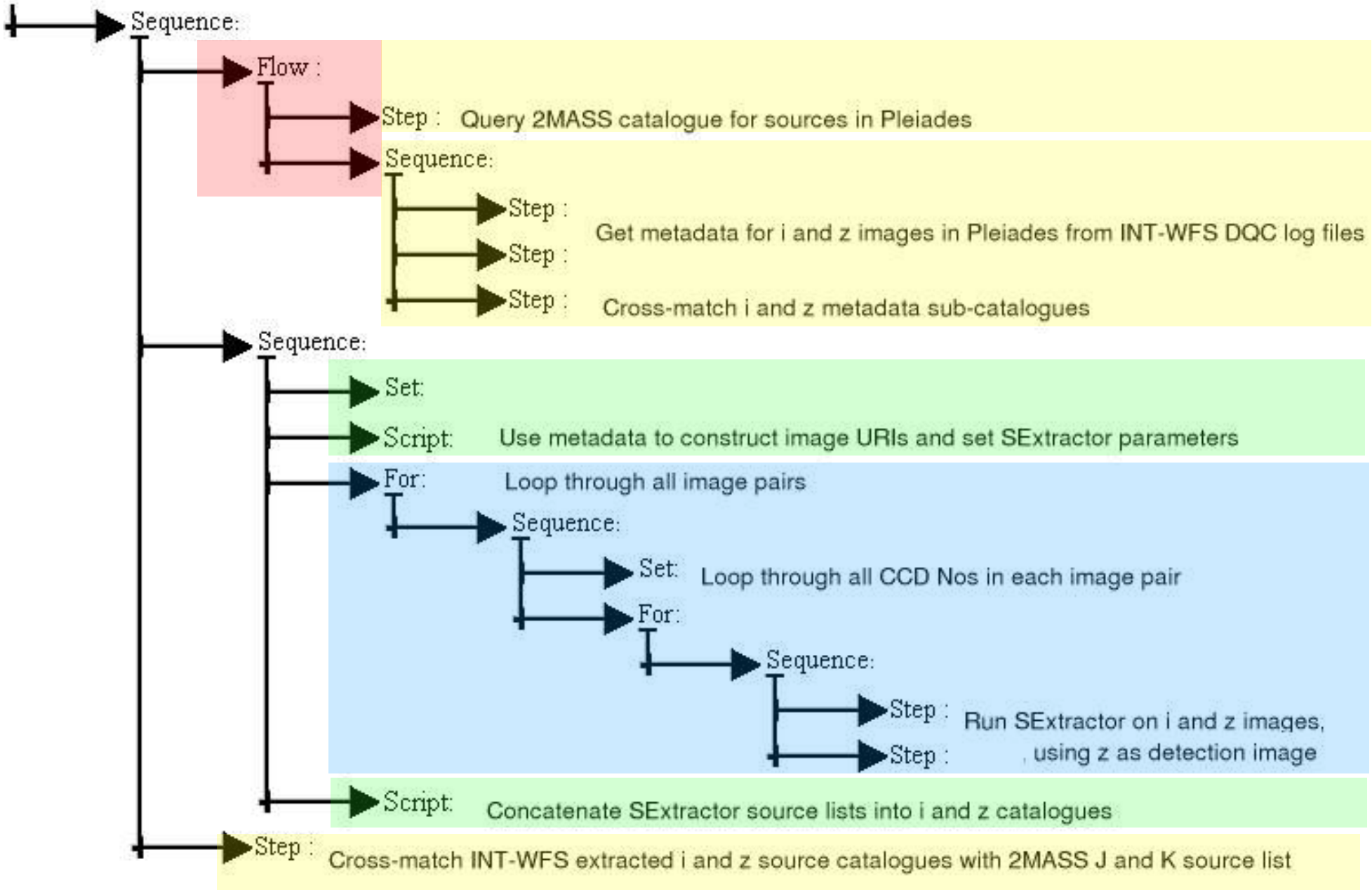
- Query 2MASS catalogue at ROE
- Query INT-WFS Log at Cambridge
- Feed images to SExtractor at JBO
- Cross-match source lists in MySpace at Leicester
- Use TopCat on MySpace files, display in Glasgow

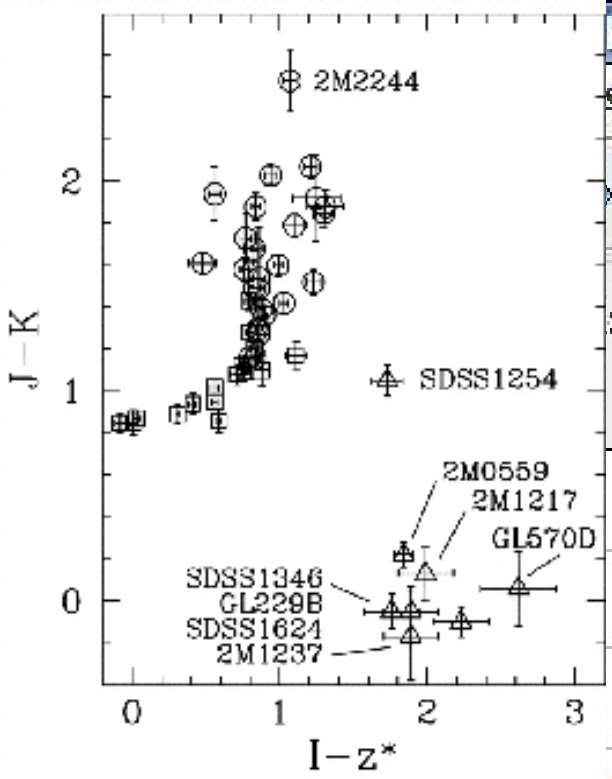


Name:	Pleiades colour cut
Description:	Pleiades colour cut using INT-WSF i and z (from images) and 2MASS J and K (catalogu

update workflow details







le Plotter

s Marker Types Points Regression Help

: 233784-5_1_izJK.vot

All

J-K / mag

i-z / mag

Overview
Introdu
Latest
FAQ
Getting
Contac
Report
About
docum
Broken
Downloa
Releas
Clients
Web S
Install
Client
Kits
Users an

Brown Dwarf summary

- >200x ~2 arcmin ($8x10^6$pixel) x *i,z* INT-WFS fields
  - ~2000 sources extracted per field (half a million possible sources)
- ~5000 2MASS sources
- 0 to 10 INT-WFS/2MASS cross-matches per field
- So far, average one L-dwarf candidate per field
- Any T-dwarfs? Still to be found.
- Future: Data mine for
  - distances,
  - proper motions,
  - Li/CH_4 lines
- *Method could also find free-floating planets (Lucas et al.)*

Lecture 2: Acknowledgements

- Brown dwarf science example – slides 31 to 39 - adapted from Anita Richards: Brown Dwarf science case developed for the AstroGrid Dec 2004 demo – see http://wiki.astrogrid.org/bin/view/Astrogrid/AgDemoDec2004Galactic
- Skyquery slide 27 adapted from Tamas Budavari: http://www.us-vo.org/summer-school/proceedings/presentations/Budavari-VoStandards.ppt
- IVOA standards – see http://www.ivoa.net/forum/

Next Lecture:
Images, Applications and Workflows