Exploiting Virtual Observatory and Information Technology: Techniques for Astronomy

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

Lecture #3 Goal:
Applications and Theory,
Workflows
Summary: Lecture #3

- Introduction
- Standards for Data Access
- Applications in a VO
  - Common Execution Architecture
  - Workflows
- Theory in a VO
- Science Example
  - Photometric redshifts
  - ... feeding into a observation/model SED case
Introduction: Image Access

• Data exists in catalogue form and image form
  – Access to catalogues discussed in Lecture 2
  – Access to images (includes theory data, spectra and so forth)

• Virtual Observatory standards to address these
  – Data Access Layer (SIA, SSA)
  – Data Models
  – Common Execution Architecture
  – Workflow
Data Access Layer: DAL
http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDAL

• Provide standards to enable access to data sets

- A dataset can be complex
  - Concept of DAL stds to allow access to dataset elements: image, spectrum, SED, etc
Simple Image Access: SIA

http://www.ivoa.net/Documents/latest/SIA.html

- Protocol for retrieving image data from a variety of astronomical image repositories through a uniform interface
- Provides access to 'image' data
  - regularly sampled (pixelated) data
  - (instead of spectrum, catalog, etc.)
  - usually an image of the sky, with a World Coordinate System (WCS)
- Service-oriented data discovery
  - query service to discover data
- Access to image metadata
  - can get image metadata without retrieving the actual image
  - uniform description based on standard data models
- Access to image datasets
  - data may be virtual or computed on demand
  - uniform interface to any type of image data
Simple Spectra Access: SSA

- Provides access to 'spectral' data
  - similar to SIA but deals with tabular spectrophotometric data
- Service-oriented data discovery
  - query service to discover data
- Access to dataset metadata
  - can get dataset metadata without retrieving actual dataset
  - uniform interface based on standard data models
- Access to actual dataset
  - data may be virtual, i.e., computed on demand
  - uniform interface to any type of spectral data
  - hides details of how data is stored or represented externally
Virtual Data

• Key VO concept: most data that is worked on is Virtual Data, i.e. Data that is created 'on the fly' during the VO process
  – Creation of virtual data may involve additional processing, e.g. Calibration of raw data from a request to an archive, subsetting of images etc.
  – Access conforms to 'Data Model' standards

• Basic SIAP usage:
  – Position and Size of search region sent via an HTTP GET call
  – Query response is a VOTable describing the images
    • Also gives an access reference to the actual data
  – Get the data via a fetch – use the returned file refs to actual data:
    • e.g http://cass38.ast.cam.ac.uk/cgi-bin/wfs-siap/getImage?run=347497&ccd=2
    • The reference URL may point to a service which returns a processed image, e.g. A cutout, a mosaic.
SIA and SSA Services

- A wide range of SIAP (the 'P' is for protocol) and SSAP services now provided to give access to image data
  - Cutout services
  - Mosaic services
  - Pointed observations
- Services discoverable through the 'Registry'
- NVO 'datascope' actions multiple SIAP/SSAP calls
  - Demonstrated in Lect 5
  - Science case later...
Data Models:
http://www.ivoa.net/twiki/bin/view/IVOA/IvoaDataModel

- Define standards to describe the structure and semantics of astronomical data
  - Generic data model broken down into key elements:
    - High level: image, spectrum, time, catalog
    - Low level: quantity, resolution
  - Provenance
    - IDHA model used by CDS
    - Specific SED example
    - DM defines other standard VO interfaces
IDHA Data Model Example:
http://alinda.u-strasbg.fr/IDHA/lastmodel/
SED: Spectral Energy Distribution:
http://hea-www.harvard.edu/~jcm/vo/docs/spec0.92.html

Model to enable construction of a multiwavelength SED from spectral AND photometric data points
Applications @ IVOA:
http://www.ivoa.net/twiki/bin/view/IVOA/IvoaApplications

- Applications to manipulate and process data
  - Client side tools: conforming VO standards: e.g. Aladin
  - Server side tools: e.g. 'hyperZ' running on 1000's of images

- Applications now being developed to exploit standard interfaces
  - Those employing VOTable for data exchange
    - VOPlot, TopCat, Mirage, Aladin
  - Large scale service applications, e.g. Montage

- For the server side systems, concept of a framework to allow the use of a wide variety of applications in user configurable workflows.

- More on specific applications in Lect 5 + 6.
Common Execution Architecture

• Models how an application is run in the Virtual Observatory
  - An application is any process that consumes or produces data

• Set of interface definitions and schema
  - Defines the tool (application) and its parameters
  - How to execute the tool
    • Initialise
    • Registers listeners for logging and results
    • Gather remote data
    • Actually runs the application

• See http://www.astrogrid.org/maven/docs/HEAD/applications/
Workflow

- Work is run remotely and asynchronously
- Archives searched and results manipulated
- Results are stored in a virtual file system
- Queries and workflows can be re-used and shared
Workflow: continued

- Workflow enables complex operations to be carried out, with the exact details of where the operations are occurring being hidden from the user
  - Gives access to CPU away from the desktop
- Workflows can be shared, amended – enables community sharing of processes in addition to results
- AstroGrid is currently the only VO project with a workflow workbench where scientific workflows can be created and run.
  - This workflow engine is being integrated into the Euro-VO
- Workflows are constructed via discovery of relevant data and applications from the Registry
  - Applications are provided through the CEA
Theory in the VO

• A wide problem:
  – Large scale: e.g. Stellar hydro-codes, N-body simulations
  – Small scale: e.g. Spectral synthesis codes
  – Fundamental data: e.g. Excitation rate coefficients

• In principle treat simulation data as observational data
  – Enable comparison of simulations with observations
  – Enable comparison of simulations with simulations

• Early effort in areas of Data Models, Metadata
Virgo simulations

- Accessing observable properties computed from the simulations
Science Example

Putting the technology to use ...
Science Case: Galaxy Distances
http://wiki.astrogrid.org/bin/view/Astrogrid/AgDemoDec2004Extragalactic

- Determine the distance to galaxies
  - Use of broad band photometry is efficient for large samples
  - Relies on the identification of spectral breaks in galaxies' spectral energy distribution

- Aim here is to automate the generation of the galaxy fluxes from survey image data, and feeding these fluxes into a number of specialist applications which return statistical estimates of galaxy redshifts (and thus distance)
Extragalactic Case Workflow

Image Selection (WFS image data)

SIAP Image call

U image \(\rightarrow\) U cat

\(g\) image \(\rightarrow\) g cat

r image \(\rightarrow\) r cat

i image \(\rightarrow\) i cat

Z image \(\rightarrow\) Z cat

UgriZ band Federation

UgriZ merged catalogue

HyperZ Photometric redshift

redshift hyperz

Bpz Photometric redshift

redshift bpz

redshift Federation

objects and redshifts
Accessing the image data

- INT Wide Field Survey data
- Data obtained using the Wide Field Camera on the INT
  - Accessed through an IVOA std SIAP call
  - Returns lists of files
  - These uploaded to MySpace via a 'workflow script'
- The data is held in Cambridge at CASU
- The MySpace server is in Leicester
Generating the image catalogues

• For each image field pointing, the mosaiced image data is returned on the basis of individual CCD images
  – One INT WFC pointing is a mosaic image with 4 CCD images
  – Each CCD image 16 MB: one pointing in 5 colours > 1/3 GB

• Each field pointing observed in 5 colours

• Each image file processed with Sextractor
  – This application returns positions of objects and fluxes

• Outputs for each colour federated
  – Metadata added to output file
  – This uses a VO federator application

• Sextractor runs at Jodrell

• The Federator runs in Leicester
The Workflow: Sextractor

Rich scripting capability
Hyperz and Bpz: The redshift apps

• The object photometry catalogues are fed into two apps
  – Hyperz: determines redshifts by template fitting to the SEDs
  – Bpz: Bayesian technique, similar to hyperz but includes weighting to reduce degenerate fits at different redshifts

• Applications run at Jodrell
  – Both compute intensive, generating large hyper cubes

• End step is the creation of two output files > then merged

• Final output catalogue:
  – Per field contains objects, photometry and redshifts
  – Integration with external viewers such as AVO/Aladin
Optical – IR – Redshifts
Example Usage: 1st Science from the AVO

- Search for high luminosity Type 2 QSO's – Optically obscured due to viewing thru the dust torus
- Use of deep multicolour data – GOODS – HST ACS, XMM/Chandra X-Ray
- Use of AVO demonstrator tool – Revealed 4 times more Type-2 quasars in sample – Tool demonstrates ease of access to data
  - Enabled via interoperability standards – Workflow enabled mass scale redshifts
Science Case Extension

Extending this science case ...
... this brings many elements of VO technologies lectures 1, 2, and 3 together ...
Observation/ Model Comparisons

http://www.euro-vo.org/twiki/bin/view/Avo/AvoDemo2005Gal

• Observational Data:
  – At a position, return broadband photometric catalogues
  – Manipulate to generate uniform Spectral Energy Distributions
  – Plot and study the SEDs
  – Select samples: e.g. based on colours

• Theoretical Data:
  – For multiple spectral synthesis models
  – Generate multiple theoretical SEDs (e.g. For range of ages)

• Comparison
  – Access and compare/fit observational and model SEDs

• Iterate
  – Home in on parameter space to generate 'best fit' models
Euro-VO Portal

- User login
  - Authentication
  - Authorisation
  - Communities
- Credentials in Registry
- Central Access
  - All further components
- Portal runs in Garching
  - Registry in Leicester
  - Data Access in Cambridge, Edinburgh, Manchester, Strasbourg
  - Applications in Garching, Cambridge
  - Thus one user runs processes over a distributed network.
Querying the catalogue data

- ELIAS photometric catalogue
  - Held in queriable database

- Query constructed with IVOA std ADQL
  - Table metadata uploaded from the registry call
  - Query saved to MySpace

- Workflow element
  - Send data query to the database (CDS)

- Workflow execution
  - Query sent through a standard VOQL call to the database
  - Results of query returned to MySpace in a std VOTable file
Workflow: Generating the SEDs

• The ELAIS catalogue contains 3500 objects
  - IR + opt photometry

• Problem:
  - IR fluxes, but optical magnitudes

• Solution:
  - Convert magnitudes to fluxes based on magnitude zero points

• Script
  - Takes input catalogue
  - Converts magnitudes to fluxes, and creates per object SED files
  - Each SED file in a standard format
  - Script implemented in Groovy (http://www.groovy.org)

• Workflow:
  - Find catalogue – retrieve catalogue – generate SEDs
Visualising the SEDs

- SEDs stored in MySpace
  - SED Index: Simple Spectrum Access standard
- Accessible:
  - The portal
  - External applications: e.g. Topcat, Treeview
  - In this case Aladin and Specview
- Aladin:
  - MySpace browser
  - Select SED
  - Each SED displayed in an Aladin plane
  - In the metadata browser, click 'view' spectrum
    - SED transferred to Specview for analysis
SED generate & visualise
Pegase and GALEXV: The spectral synthesis model applications

- Two techniques to generate theoretical galaxy spectral energy distributions.
- Search 'registry' to 'discover' relevant applications.
- For each, include in workflow:
  - User ability to alter input parameters
  - Each run generates sets of output spectra
  - Output formats conform the VO interoperability standards
- Applications run within the AstroGrid CEA
  - Common Execution Architecture
    - Standard framework for applications
    - Metadata describes application, discovererable through 'Registry'
Observation/ Model Comparisons

• Observational and Theory data stored in MySpace
  – 'one click' access through Aladin
  – Read SEDs into Specview
  – Read models into Specview

• For any observed SED
  – Select and 'double left click'
  – Allows overplotting of model
  – Select best model
    • Rerun models to tune input parameters for best match

• Automatic model/observed SED fitting
  – This tool required – possible development through VOTECH.
SED fitting ...
Lecture 3: Acknowledgements + Refs

- SSA slides 5 to 7 adapted from Doug Tody: see http://www.us-vo.org/summer-school/proceedings/presentations/dal-nvoss.ppt
- IVOA standards – see http://www.ivoa.net/forum/
- Pegase: see http://www2.iap.fr/users/fioc/PEGASE.html
- Galexv: see http://www.cida.ve/~bruzual/bc2003
Next Lecture:
Mining the Sloan Digital Sky Survey