Virtual Observatory

Jaroslav Vazny

Department of Theoretical Physics and Astrophysics

Masarykova Univerzita

A thesis submitted for the degree of

Master

Yet to be decided

I would like to dedicate this thesis to my loving parents ...

Acknowledgements

And I would like to acknowledge ...

Abstract

This is where you write your abstract ...

Contents

List of Figures

Introduction

From the dawn of existence astronomy has always been starved for data, but in the last few decades the situation has changed and now we are facing the data flood of biblical proportions. The data are not just increasing in size but in complexity and dimensionality. Ball and Schade [2010] Astroinformatics is the new field of science which has emerged from this technology driven progress. Virtual Observatory, Machine learning, Data Mining, Grid computing are just few examples of new tools available to scientist.

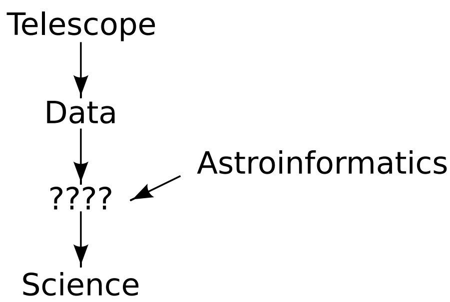

Figure 1: Astroinformatics in the context of astronomy Ball and Schade [2010]

Of course astronomers are not alone and particle physics, biology and other sciences are also in the vanguard of the data intensive science. This is great opportunity for interdisciplinary collaboration.

This work deals with the problem of semi-automatic procedures for finding Be stars Porter and Rivinius [2003] candidates in the astronomy surveys. More than straight forward process it's trail and error approach probing new possibilities with rather interesting that useful results.

The aim of this work is to be introductory to the technologies of Virtual Observatory and Data Mining and for this reason it is intended to have following properties:

- Main Chapters starts with questions answered in the text and diagram to ease orietation,

- is full of examples,
- is non-linear in nature,
- is meant to be compact and consistent,
- is far from complete.

Chapter one is an introduction to the technologies related to Virtual Observatory. The motivation behind the concept is given without paying too much attention to historical details. Main principles and protocols are discussed and explained. Important aspect are demonstrated on numerous examples. Chapter two is an introduction to Machine Learning and Data Mining in the context of astrophysics. Only methods used in practical part of this work are described in detail: Decision Trees and Support Vector Machines. Examples of several classifications are demonstrated. Third chapter introduces problematic of Be stars. Chapter Four is practical application of previously described technologies and methods. Training data of confirmed Be stars from Ondrejov are correlated with others catalogues to obtain color indexes and spectra. Results are processed by Data Mining algorithms using several libraries and tools. In the last chapter achieved results are critically discussed.

Activities related to this work go beyond this text. Wiki pages[1] were created to present the results and discuss related topic with supervisor as well as with others scientist around the world. Several programs were created to analyze and process acquired data. Source codes were maintained by GIT version system allowing easy sharing. All software used and produced are open source.

[1] http://physics.muni.cz/~vazny/wiki/index.php/Diploma_work

Chapter 1

Virtual Observatory (VO)

What is the motivation behind Virtual Observatory? Is data avalanche problem only in astronomy? What is IVOA? What is Virtual Observatory architecture?

1.1 Data avalanche: Opportunity or disaster?

There are two important trends in current astronomy surveys:

- Size: The cumulative compressed data holdings of the ESO archive will reach 1 PetaByte by 2012 Hanisch and Quinn [2010]. Projects like Large Synoptic Survey Telescope (LSST) will produce about 30 TB per night, leading to a total database over the ten years of operations of 60 PB for the raw data Becla et al. [2006].
- Complexity: Modern surveys will cover the sky in different wave-bands, from gamma- and X-rays, optical, infrared, through to radio. The ability to cross correlate these observations together may lead to the new understanding of physical phenomenas. Hanisch and Quinn [2010]

Such amount of data is not possible to transfer over the network. Data resources are heterogeneous, distributed and decentralized in nature.

There is an interesting analogy with the problem (and the solution) which had scientist during LEP project at CERN. Their problem was too many documents in different formats. Tim Berners-Lee[1] designed set of protocols (URIs, HTTP and HTML) which allowed link and share documentsBerners-Lee and Cailliau [1990]. This was recognized as generally useful and World Wide Web was born. An important role plays the World Wide Web Consortium (W3C) in developing Web standards[2].

[1] Sir Timothy John "Tim" Berners-Lee. British engineer and computer scientist and MIT professor credited with inventing the World Wide Web.

[2] Prior to its creation, incompatible versions of HTML were offered by different vendors, increasing the potential for inconsistency between web pages.

1.2 International Virtual Observatory Alliance (IVOA)

Figure 1.1: IVOA members

What is necessary is sets of standards and protocols to deal with heterogeneous distributed data and the authority which encourages their implementation. Such authority is the International Virtual Observatory Alliance (IVOA). It comprises 19 VO programs from Argentina, Armenia, Australia, Brazil, Canada, China, Europe, France, Germany, Hungary, India, Italy, Japan, Russia, Spain, the United Kingdom, and the United States and intergovernmental organizations (ESA and ESO)Hanisch and Quinn [2010].

Standards and specifications produced by IVOA can be obtained at http://www.ivoa.net/.

1.3 Architecture

The Architecture is depicted on the figure 1.2.The level of abstraction goes from top to bottom. Starting with interfaces, used by people or applications to discover resources. Next level is the service layer implemented by standard protocols, followed by the hardware level where actual data are stored. This onion like structure hide the complexity of the lower layer and provide data and meta-data to the higher layer. This concept is similar to TCP/IP [1] protocol.

[1]TCP/IP (Transmission Control Protocol/Internet Protocol). The basic communication language or protocol of the Internet.

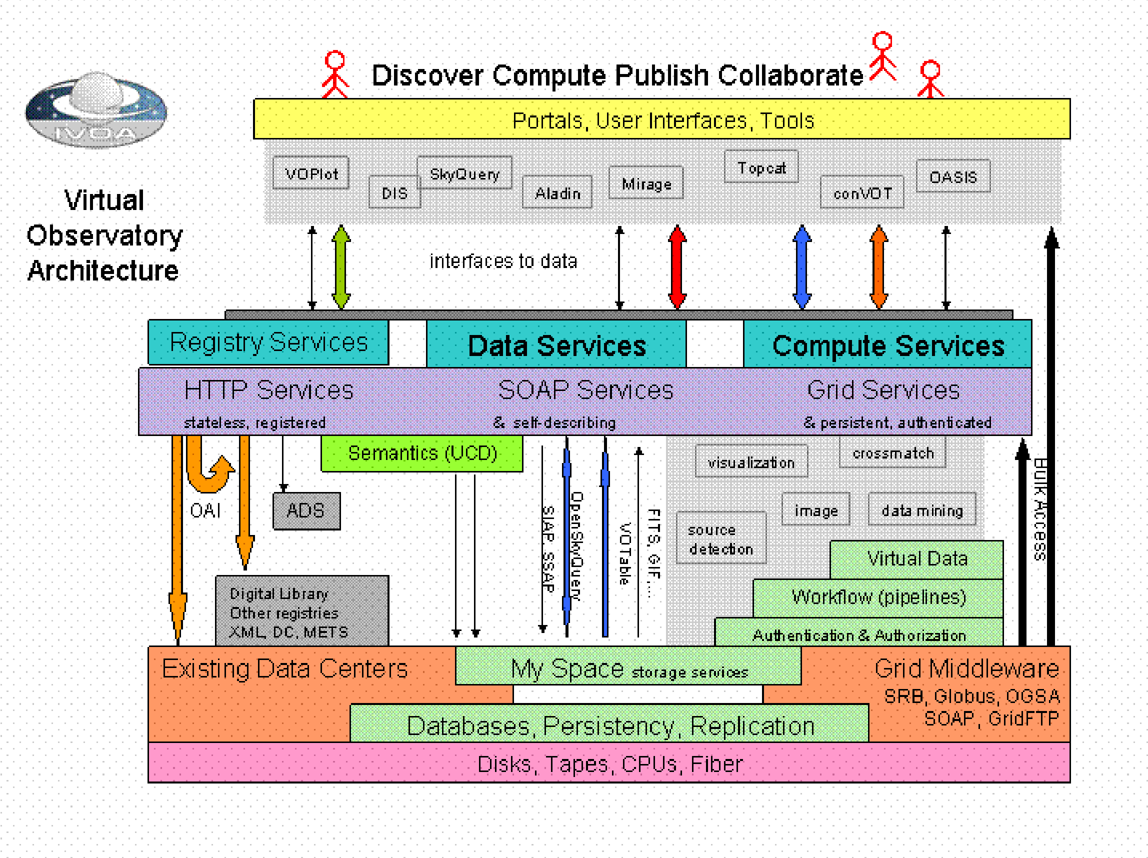

Figure 1.2: VO Architecture

The essence of VO architecture is service orientation. Each service is autonomous with well defined defined boundaries. Very important aspect of VO implementation is the adoption of formats and protocols used in astronomy (FITS) and computers science (XML [1] , Web service [2] SOAP [3]) for many years. In other words VO does not try to reinvent the wheel but it's stands on the shoulders of giants.

1.4 VOResources

A resource is a general term referring to a VO element that can be described in terms of who curates or maintains it and which can be given a name and a unique identifier. Just about anything can be a resource: it can be an abstract idea, such as sky coverage or an instrumental setup, or it can be fairly concrete, like an organization or a data collection. Benson et al. [2009]

UML[4] diagram of the resource in on the figure 1.3. Next paragraph is an attempt to explain thsi diagram to non-programmers. Full arrow means generalization, Resource can be a generalization of organization, data collection, application or service. Single arrow means association. Organization can be linked (assosiated) together with other organization (multiplicity is represented by number 1, 0..). The same is true for data collection. Organization is a generalization of and/or provider which can own zero to N services. Diamond means aggregation Publisher can have any resources.

[1] Extensible Markup Language (XML) is a set of rules for encoding documents in machine-readable form.

[2] method of communication between two electronic devices over a network.

[3] Simple Object Access Protocol, is a protocol specification for exchanging structured information in the implementation of Web Services in computer networks.

[4] Unified Modeling Language. Standardized general-purpose modeling language in the field of object-oriented software engineering.

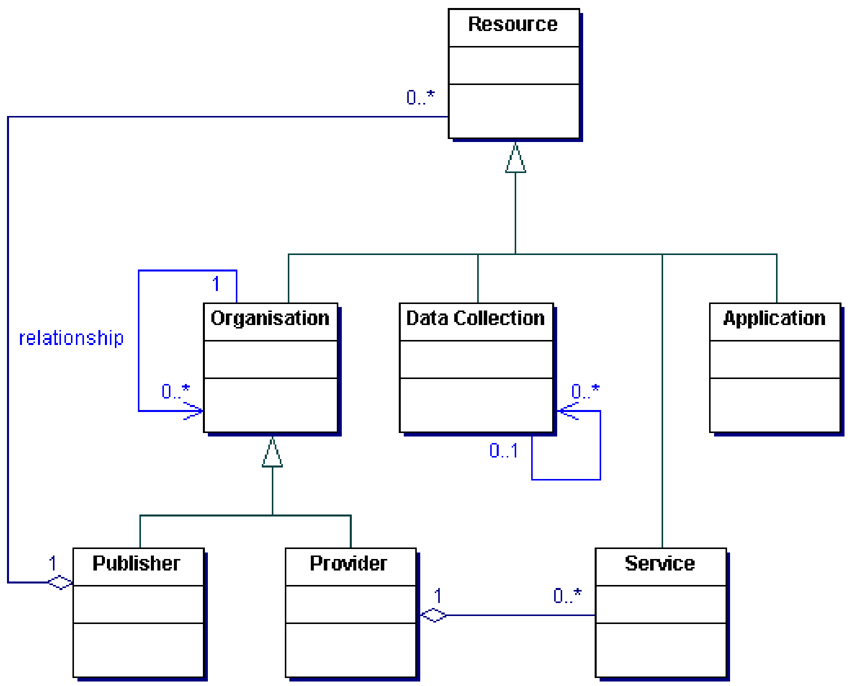

Figure 1.3: UML diagram of VOResource

Following example uses program stilts[1] to query registry with parameter shortName equal to 'AIASCR'[2]. This return VOTable containing informations about the resource.

```
stilts regquery query="shortName like 'AIASCR'"
regurl=http://registry.euro-vo.org/services/RegistrySearch
ofmt=votable-tabledata > resourceExample.vot
```

Rows 1–4 define XML nad VOTable schema with adequate locations (xmlns[3]) followed by informations about the actual resource. The listing is abbreviated.

```
1  <?xml version='1.0'?>
2  <VOTABLE version="1.1"
3   xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
4   xmlns="http://www.ivoa.net/xml/VOTable/v1.1">
5  .
6  <DATA>
7  <TABLEDATA>
8    <TR>
9      <TD>ivo://asu.cas.cz</TD>
10     <TD>AIASCR</TD>
11     <TD>Astronomical Institute of the Academy of Sciences of the Czech Republic
           Naming Authority</TD>
12     <TD>http://stelweb.asu.cas.cz/web/index/index-en.php</TD>
13     <TD>Petr Skoda &lt;skoda@sunstel.asu.cas.cz&gt;</TD>
```

1.5 Data Access Protocols

1.5.1 Cone Search Protocol

1.5.2 Simple Image Access Protocol

1.5.3 Simple Spectra Access Protocol

1.6 Data Formats

1.6.1 VOTable

Motivation

VOTable is flexible storage and exchange format fundamentaly interconnected with Virtual Observatory. It has features for big-data and Grid computing. Data can be stored

[1]STIL Tool Set. Set of command-line tools based on STIL, the Starlink Tables Infrastructure Library.
[2]Astronomical Institute of the Academy of Sciences of the Czech Republic
[3]XML namespaces. Provide uniquely named elements and attributes in an XML document.

in different ways in dependence on the charatecter and size. Small tables can be stored in pure XML [1], while large-scale data can be referenced with the URL [2] syntax protocol://location. It combine web standards (it is based on XML) and astronomy tradition in storing data (it is FITS compatible). Expiration and authentication are also supported.

Structure

Following example of VOTable was created from SDSS FITS file used in this work. First there is an information about XML and VOTable versions and references to corresponding XML Schema [3]. ¡TABLE¿ tag encapsulating tabular data. ¡FIELD¿ tag describe identifinaction (ID), type and precision of columns. ¡DATA¿ tag contains data (here) in TABLEDATA format (other types are FITS and BINARY)

```
<?xml version="1.0" encoding="utf-8"?>
<!-- Produced with vo.table version 0.6
     http://www.stsci.edu/trac/ssb/astrolib
     Author: Michael Droettboom <support@stsci.edu> -->
<VOTABLE version="1.0"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:noNamespaceSchemaLocation="http://www.ivoa.net/xml/VOTable/v1.0"
 xmlns="http://www.ivoa.net/xml/VOTable/v1.0">
 <RESOURCE type="results" >
  <TABLE >
   <FIELD ID="col0" name="wave" datatype="float" unit=""
   precision="F9"/>
  <DATA>
    <TABLEDATA>
     <TR>
      <TD>4012.50757</TD>
     </TR>
 </TABLEDATA>
   </DATA>
  </TABLE>
 </RESOURCE>
</VOTABLE>
```

[1] Extensible Markup Language. W3C standard. Set of rules for encoding documents in machine-readable form

[2] Uniform Resource Locator. Uniform Resource Identifier (URI) that specifies where an identified resource is available and the mechanism for retrieving it.

[3] Define the legal building blocks of an XML document. Note: XML schema can be described by XML Schema, Document Type Definition (DTD) or RELAX NG

Tools and libraries

Examples

All example created using ATpy[1]

Following example shows transformation FITS into VOTable.

```
In [1]: import atpy
In [2]: tbl = atpy.Table('spSpec-53401-2052-458.fit',hdu=1)
Auto-detected input type: fits
In [3]: tbl.write('votableExample.xml')
Auto-detected input type: vo
```

1.6.2 FITS

Motivation

"An archival format must be utterly portable and self-describing, on the assumption that, apart from the transcription device, neither the software nor the hardware that wrote the data will be available when the data are read." Council [1995]

FITS (Flexible Image Transport System) was originally created for data exchange between WSRT [2] and the VLA [3] Schlesinger [1997]. It is now used as a file format to store, transmit, and manipulate scientific data and it is (thanks to it's revolutionary design) de facto standard in astronomy.

Structure

One file can contain several HDUs (Header Data Units).The first part of each HDU is the header, composed of ASCII card images containing keyword=value statements that describe the size, format, and structure of the data that follow.

- Primary header and data unit (HDU).
- Conforming Extensions (optional).
- Other special records (optional, restricted).

Standards and documents related to FITS are maintained by IAUFWG [4] and aviable at http://fits.gsfc.nasa.gov.

[1] High-level Python package providing a way to manipulate tables of astronomical data in a uniform way.

[2] Westerbork Synthesis Radio Telescope

[3] Very Large Array

[4] International Astronomical Union FITS

Examples

There are many libraries for working with FITS files. The official list is aviable at http://fits.gsfc.nasa.gov/fits_libraries.html. PyFITS, library for Python programming language was used for following examples. PyFITS is a development project of the Science Software Branch at the Space Telescope Science Institute http://www.stsci.edu/resources/software_hardware/pyfits.

Reading FITS headers.

```
In [1]: import pyfits
In [2]: hdulist = pyfits.open('spSpec-53237-1886-248.fit')
In [3]: hdulist.info()
Filename: spSpec-53237-1886-248.fit
No.    Name         Type      Cards   Dimensions   Format
0    PRIMARY    PrimaryHDU    213  (3874, 5)    float32
1               BinTableHDU    54  6R x 23C     [1E, 1E, ...
2               BinTableHDU    54  44R x 23C    [1E, 1E, ...
3               BinTableHDU    18  1R x 5C      [1E, 1E, ...
4               BinTableHDU    32  53R x 12C    [1J, 1J, ...
5               BinTableHDU    26  36R x 9C     [19A, 1E, ...
6               BinTableHDU    14  3874R x 3C   [1J, 1J, 1E]
```

Printing primary HDU.

```
In [4]: print hdulist[0].header
-------> print(hdulist[0].header)
DATE-OBS= '2004-08-20'         / 1st row - TAI date
TAIHMS  = '10:36:18.11'        / 1st row - TAI time (HH:MM:SS.SS) (TAI-UT = appr
TIMESYS = 'tai     '           / TAI, not UTC
TAI-BEG =        4599713999.00 / Exposure Start Time
TAI-END =        4599717089.00 / Exposure End Time
MJD     =                53237 / MJD of observation
MJDLIST = '53237 '             /
VERSION = 'v3_140_0'           / version of IOP
CAMVER  = 'SPEC1 v4_8'         / Camera code version
OBSERVER= 'prn     '
OBSCOMM = 'science '
TELESCOP= 'SDSS 2.5-M'         / Sloan Digital Sky Survey
```

Updating FITS file.

```
  In [16]: prihdr = hdulist[0].header
  In [17]: prihdr.update('observer', 'Astar')
  In [18]: prihdr.add_history('I updated this file 3/27/11')
```

Example from program pf (plot fits) created for purposes of this work to plot $H\alpha$ emission in the spectra.

```
def read(file):
   """ Read fits file. Convert wavelength to angstroms """
   data = pyfits.getdata(file)
   w = lambda x : 10.0**(3.5796 + x*10.0**(-4))
   x = np.arange(1,data[0].size + 1)
   xx  = w(x) # convert to actual wavelenght
   return np.asarray([xx, data[0]])

def plot(file,xdata,ydata,spLine):
   fig = plt.figure()
   ax = fig.add_subplot(111)
   graph = ax.plot(xdata,ydata, 'r')
   ax.set_title(file)
   ax.set_xlabel("$Wavelenght [\\AA]$")
   ax.set_ylabel("$Energy [10^{-17} erg/s/cm^2/\\AA]$")
   ax.axvline(x=spLine, color = 'g', ls ='--')
```

I would also like to add an extra bookmark in acroread like so ...

References

N.M. Ball and D. Schade. ASTROINFORMATICS IN CANADA. 2010. v, 1

J. Becla, A. Hanushevsky, S. Nikolaev, G. Abdulla, A. Szalay, M. Nieto-Santisteban, A. Thakar, and J. Gray. Designing a multi-petabyte database for LSST. *Arxiv preprint cs/0604112*, 2006. 3

K. Benson, R. Plante, E. Auden, et al. IVOA Registry Interfaces. *IVOA Working Draft*, 2009. 6

T. Berners-Lee and R. Cailliau. WorldWideWeb: Proposal for a HyperText project. *European Particle Physics Laboratory (CERN)*, 1990. 3

National Research Council. Preserving Scientic Data on our Physical Universe. 1995. 10

RJ Hanisch and PJ Quinn. The international virtual observatory. *Retrieved from http://www. ivoa. net/pub/info/TheIVOA. pdf on*, 24, 2010. 3, 4

J.M. Porter and T. Rivinius. Classical Be Stars. *Publications of the Astronomical Society of the Pacific*, 115(812):1153–1170, 2003. ISSN 0004-6280. 1

B. Schlesinger. A Users Guide for the Flexible Image Transport System. 1997. 10