# 1 Introduction

Some text ...

# 2 The IVOA Virtual Observatory

The International Virtual Observatory Alliance (IVOA) [1] was formed in June 2002 with a mission to

> "facilitate the international coordination and collaboration necessary for the development and deployment of the tools, systems and organizational structures necessary to enable the international utilization of astronomical archives as an integrated and interoperating virtual observatory."

The work of the IVOA focuses on the development of standards, providing a forum for members to debate and agrees the technical standards that are needed to make the VO possible.

The Virtual Observatory (VO) is the realization of the IVOA vision of an integrated and interoperating virtual observatory.

The operational VO itself is comprised of a global shared metadata registry, along with individual data discovery and data access services deployed at each of the participating institutes, which work together to present a uniform mechanism for discovering and accessing data, irrespective of where it is physically located.

The VO architecture and data discovery process is very similar to the interconnected metadata collections approach described in a paper on 'The new bioinformatics: integrating ecological data from the gene to the biosphere' (Jones et al. 2006). [2] [3] [4]

> "An alternative, more robust approach to the highly structured, vertically integrated data warehouse is a more loosely structured collection of project-specific data sets accompanied by structured metadata about each of the data sets."

> "Each of the data sets is stored in a manner that is opaque to the data system in that the data themselves cannot be directly queried; rather, the structured metadata describing the data is queried in

order to locate data sets of interest. After data sets of interest are located, more detailed information (such as the detailed data model that specifies, e.g., the definitions of the variables) can be extracted from the metadata and used to load, query, and manipulate individual data sets."

## 2.1 Example use case

A useful way to illustrate how the VO data discovery process works is to look at an example task such as selecting images covering a particular region of the sky, in a particular wavelength e.g. infrared, visible light, radio or xray,

### 2.1.1 Service discovery

The first step of the process is to identify the services that provide access to the type of data we are looking for by querying the VO Registry.

The VO Registry is comprised of a number of small local registries, typically hosted at the participating institute level, working in cooperation with a set of higer level global registries typically hosted by a few key institutes that aggregate the data from the smaller registries to create a global searchable index of metadata about all the services and datasets in the VO.

When a new service is deployed, part of the deployment process involves registering the service with the local registry. The local registry is then responsible for collecting and storing the metadata that describes both the service itself and the datasets that it provides access to.

Once the metadata is registered in a local registry, it is automatically propagated up to the next level and replicated between the global registries.

This means that a client can access all of the available metadata for all of the services and datasets in the VO by querying any of the global registries.

The first step in fulfilling our example use case is to identify services that contain the type of data we are looking for, in this case images, by querying the registry for services that support the Simple Image Access (SIA) [5] capability.

In addition to the technical details of services and their capabilities the regitry metadata also contains details about the content of datasets, including details of the wavelength(s) measured, e.g. infrared, visible, radio or xray.

This allows us to refine our query to search for SIA services that contain images in a specific waveband, e.g. `optical` or `infrared`

The registry query would return a table of data, each row of which contains information about a SIA services that contains the type of data we are interested in - images in a particular wavelength.

The VO is itself an evolving system, building on the existing work to add additional levels of integration as new features are added to the IVOA specifications.

A recent addition to the IVOA is the HEALPix Multi-Order Coverage Map (MOC) [6] which will allow registry services to perform coarse grained region matches.

Once this is in place we should be able to further refine our query to filter for SIA services that contained data in a particular region of the sky.

### 2.1.2 Data discovery

The next stage of the process is to query each of the SIA services in the list to discover details about the individual images available from that service.

A SIA query can specify parameters for a particular wavelength and a particular region of the sky :

- `POS` The positional region
- `BAND` The the energy interval

Each SIA service would return a table of data, each row of which contains metadata about an individual image. The details of the fields in the image metadata are defined in the ObsCore [7] data model.

As every one of the SIA services returns a standard response, it makes it easy to combine them to produce a single list of all the images available in the VO that match our search criteria.

The user can then select which data products they are interested in, and their client software can use the metadata in the SIA results to access the individual data products and display them in the appropriate tools.

The two key components of this process are

- A standard interface for the global registry that uses a standard set of attributes to describe datasets and services
- A standard interface for local data access services that uses a standard set of attributes to describe the available data products

The separation between the initial service discovery query at the global level followed by individual data discovery queries at the local level are very similar to the stages described in (Jones et al. 2006) of first querying the metadata to establish the data location followed by a more detailed individual queries to establish what the data is and how to access it.

## 2.2 Carbon density comparison

We can compare the VO data discovery process for astronomy data with an example use case based on a recent study 'Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites' (Mitchard et al. 2014) [8] [9] comparing remote sensing data from satelites with ground plot data collected in the field.

The study compares two sets of remote sensing data, from the NASA Jet Propulsion Laboratory (Saatchi et al., 2011) [10] [11] and the Woods Hole Research Center (Baccini et al., 2012) [12] [13] with four sets of ground plot data from Red Amazónica de Inventarios Forestales (RAINFOR) (Malhi et al. 2002) [14] the Amazon Tree Diversity Network (ATDN) (ter Steege et al., 2003) [15] the Tropical Ecology Assessment and Monitoring (TEAM) [16] network and the Brazilian Program for Biodiversity Research (PPBio). [17] [18]

### 2.2.1 Satelite data

### 2.2.2 Ground plot data

In order to calculate a single above ground biomass (AGB) dataset, the ground plot data were brought together in the ForestPlots.Net (Lopez-Gonzalez et al. 2009, 2011) [19] [20] [21] database.

ForestPlots.Net is a website and database designed to provide a repository for long-term intact tropical forest inventory plots, where trees within an area are individually identified, measured and tracked through time.

In addition to the raw measurements of tree diameter, the ForestPlots.Net database stores a comprehensive set of metadata including taxonomic information about the individual trees and detailed metadata about the plots.

Of the three sets of ground plot data, the data from RAINFOR and ATDN were already available in the ForestPlots.Net database. The plot data from the TEAM and PPBio projects were manually downloaded and imported into the ForestPlots.Net database.

The principal AGB dataset was calculated using a tropical forest model described in Chave et al. (2005), [22] [23] [24] using one of the built-in SQL queries provided by the ForestPlots.Net database system.

The resulting data set was itself stored in the ForestPlots.Net database as a new dataset available for download as part of the source material for the paper.

## 2.3 Diverse metadata formats

Within just the datasets used by our example use case, we can see a variety of different database systems storing different types of metadata in a varierty of different structures and formats.

### 2.3.1 GIVD

ForestPlots.Net is itself part of a hierarchy of databases containing metadata about databases.

The Global Index of Vegetation-Plot Databases (GIVD) [25] is a database of metadata describing databases of vegetation plot data from around the world.

ForestPlots.Net is described in the GIVD database `[GIVD:00-00-001]` [26] as is the Brazilian Program for Biodiversity Research (PPBio) information system `[GIVD:SA-BR-001]` [27] and the data from the Tropical Ecology Assessment and Monitoring (TEAM) network `[GIVD:00-00-002]`. [28]

The GIVD database authors describe the project in a 2011 paper, "The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science" (Dengler et al. 2011) [29] [30] and suggest some future applications, including the idea of combining different types of data from different, distributed, databases.

> "Our longer-term vision is to develop GIVD in ways similar to Metacat (Jones et al. 2006), so that, ultimately, users who query GIVD will not only receive information on which databases contain data suitable for the intended analyses, but they will also discover other data from distributed databases, with GIVD acting as the central node."

> "By coupling species specific trait characteristics (e.g. mean plant height, specific leaf area, growth form) found in trait databases, such

> as LEDA (Kleyer et al. 2008) or TRY (http://www.trydb.org), with plot-based distribution information on those species, GIVD could support further refinement of DGVMs."

Their idea of many different databases working together to create a larger system has a lot of similarities with the distributed architectrure of data discovery and data access services working together to create the virtual observatory.

### 2.3.2 METACAT

From paper describing PPBio [31]

> To facilitate data searches, all the metadata were converted to XML, and the PPBio has installed a METACAT server to integrate with the Knowledge Network for Biocomplexity (KNB), a network which aims to assist ecological and environmental research.

> METACAT [32] [33]

> Metacat is a repository for data and metadata (documentation about data) that helps scientists find, understand and effectively use data sets they manage or that have been created by others. Thousands of data sets are currently documented in a standardized way and stored in Metacat systems, providing the scientific community with a broad range of science data that–because the data are well and consistently described–can be easily searched, compared, merged, or used in other ways.

> Metacat is a Java servlet application that runs on Linux, Mac OS, and Windows platforms in conjunction with a database, such as PostgreSQL (or Oracle), and a Web server.

> The Metacat application stores data in an XML format using Ecological Metadata Language (EML) or other metadata standards such as ISO 19139 or the FGDC Biological Data Profile.

### 2.3.3 DataONE

Metacat features include providing a Data Observation Network for Earth (DataONE) [34] node service.

The DataONE project is a collaboration among scientists, technologists, librarians, and social scientists to build a robust, interoperable, and sustainable system for preserving and accessing Earth observational data at national and global scales. Supported by the U.S. National Science Foundation, DataONE partners focus on technological, financial, and organizational sustainability approaches to building a distributed network of data repositories that are fully interoperable, even when those repositories use divergent underlying software and support different data and metadata content standards.

DataONE defines a common web-service service programming interface that allows the main software components of the DataONE system to seamlessly communicate.

## 2.4 VO services

## 2.5 The registry

The VO Registry provides the first layer of data discovery available in the virtual observatory. The individual registry services deployed at participating institutes work together to provide a shared repository for describing datasets, data access services and data processing services in a standard way.

The IVOA Registry Interfaces standard [35] defines the web service interfaces that support interactions between applications and registries as well as between the registries themselves.

The high level structure of the registry content is defined by a set of IVOA standards, including a standard format for IVOA Identifiers [36] and the basic Resource Metadata [37]

The details of the registry metadata are covered by a set of technical standards defining the detailed XML schemas for resource metadata.

- VOResource [38]
- VODataService [39]
- Simple Data Access Services [40]

## 2.6 Service registration

VOSI . . . and stuff . . .

## 2.7 Service metadata

registry metadata queries . . . and stuff . . .

## 2.8 Service footprint

registry footprint queries . . . and stuff . . .

HEALPix Multi-Order Coverage Map (MOC) [41]

## 2.9 Data access services

The VO DataAccess services can be categorised as two types of services.

A set of type specific data discovery servces which are designed to provide simple service interfaces for discovering and acessing data of a specific type.

- Simple Cone Search (SCS) [42]
- Simple Image Access (SIA) [43]
- Simple Spectral Access (SSA) [44]
- Simple Line Access (SLA) [45]

A tabular data access services for querying tabluar data using a common query derrived from SQL.

- Table Access Protocol (TAP) [46]
- Astronomy Data Query Language (ADQL) [47]

### 2.9.1 Simple Image Access

The Simple Image Access (SIA) protocol provides

> parameter based discovery of images and datacubes, querying the service(s) with a few well known kinds of queries that cover greater than 95% of use, and getting back easily parsed summary metadata about each available data product

The Simple Image Access (SIA) data discovery service provides support for the following use cases:

- find data that includes specified coordinates (e.g. for some object)
- find data in the circle with coordinate centre and radius
- find data in a range of longitude and latitude
- find data within a specified simple polygon (one region, no holes, less than half the sphere)
- find data containing a specified energy (e.g. wavelength) or in a specified range of energy values
- find data obtained at a specified time (e.g. including a time instant) or during a specified range of times
- find data obtained with specified polarization (Stokes) states
- find data within a specified range of spatial resolution
- find data within a specified range of field-of-view
- find data within a range of exposure (integration) time

The response from a successful SIA data discovery query is a VOTable containing instances of the ObsCore [48] data model.

Each row in the results corresponds to a data product that matches the search criteria and includes details of how to access the data products or how to request additional matadata.

### 2.9.2 Simple Spectral Access

The Simple Spectral Access (SSA) protocol is similar to the Simple Image Access (SIA) protocol.

The primary differences are the type of data searched for, and the set of query parameters.

> ... discover and access one dimensional spectra ... based on a general data model capable of describing most tabular spectrophotometric data, including time series and spectral energy distributions (SEDs) as well as 1-D spectra

### 2.9.3 Simple Line Access

The Simple Line Access (SLA) protocol is similar to the Simple Image Access (SIA) protocol.

The primary differences are the type of data searched for, and the set of query parameters.

> ... retrieving spectral lines coming from various Spectral Line Data Collections ... either observed or theoretical and will be typically used to identify emission or absorption features in astronomical spectra. ... makes use of the Simple Spectral Line Data Model (SSLDM) [49] to characterize spectral lines through the use of uTypes [50]

### 2.9.4 Table Access Protocol

Table Access Protocol (TAP) is a generic protocol for accessing general table data, including astronomical catalogs as well as general database tables, with support for both synchronous and asynchronous queries.

Special support is provided for spatially indexed queries using the spatial extensions in ADQL.

A multi-position query capability permits queries against an arbitrarily large list of astronomical targets, providing a simple spatial cross-matching capability.

Deploying the same standard interface and query language across multiple sites means that cross-matching queries are possible by orchestrating a distributed query across multiple TAP services.

# A SIA search parameters

The SIA search parameters include

- `POS` The positional region
- `BAND` The the energy interval
- `TIME` The the time interval
- `POL` The the polarization state

- `FOV` The the field of view
- `SPATRES` The the spatial resolution
- `EXPTIME` The the exposure time
- `COLLECTION` The name of the data collection that contains the data
- `FACILITY` The name of the facility where the data was acquired
- `INSTRUMENT` The name of the instrument with which the data was acquired
- `DPTYPE` The data type from the ObsCore [51] data model
- `CALIB` The calibration level
- `TARGET` The target name from the ObsCore [52] data model
- `TIMERES` The temporal resolution
- `SPECRP` The spectral resolving power
- `FORMAT` The data format

# Notes

[1]http://www.ivoa.net/

[2]The new bioinformatics: integrating ecological data from the gene to the biosphere (Jones et al. 2006)

[3]doi:10.1146/annurev.ecolsys.37.091305.110031

[4]http://www.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.37.091305.110031

[5]http://www.ivoa.net/documents/SIA/

[6]http://www.ivoa.net/documents/MOC/

[7]http://www.ivoa.net/documents/ObsCore/

[8]doi:10.1111/geb.12168

[9]http://onlinelibrary.wiley.com/doi/10.1111/geb.12168/abstract

[10]doi:10.1073/pnas.1019576108

[11]http://www.pnas.org/content/108/24/9899

[12]doi:10.1038/nclimate1354

[13]http://www.nature.com/nclimate/journal/v2/n3/full/nclimate1354.html

[14]http://www.rainfor.org/

[15]http://web.science.uu.nl/Amazon/ATDN/

[16]http://www.teamnetwork.org/

[17]doi:10.7809/b-e.00083

[18]http://www.biodiversity-plants.de/biodivers$_e$col/article$_m$eta.php?DOI = 10.7809/b−e.00083

[19]http://www.forestplots.net/

[20]http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2011.01312.x/abstract

[21]doi:10.1111/j.1654-1103.2011.01312.x

[22]Tree allometry and improved estimation of carbon stocks and balance in tropical forests

[23]http://link.springer.com/article/10.1007/s00442-005-0100-x

[24]doi:10.1007/s00442-005-0100-x

[25]http://www.givd.info/

[26]http://www.givd.info/ID/00-00-001

[27]http://www.givd.info/ID/SA-BR-001

[28]http://www.givd.info/ID/00-00-002

[29]doi:10.1111/j.1654-1103.2011.01265.x

[30]http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2011.01265.x/abstract

[31]http://www.biodiversity-plants.de/biodivers$_e$col/article$_m$eta.php?DOI = 10.7809/b−e.00083

[32]http://knb.ecoinformatics.org/software/metacat

[33]https://knb.ecoinformatics.org/knb/docs/intro.html

[34]https://www.dataone.org/

[35]http://www.ivoa.net/documents/RegistryInterface/

[36]http://www.ivoa.net/documents/latest/IDs.html

[37]http://www.ivoa.net/Documents/latest/RM.html

[38]http://www.ivoa.net/documents/latest/VOResource.html

[39]http://www.ivoa.net/documents/VODataService/

[40]http://www.ivoa.net/documents/SimpleDALRegExt/20131005/

[41]http://www.ivoa.net/documents/MOC/

[42]http://www.ivoa.net/documents/latest/ConeSearch.html

[43]http://www.ivoa.net/documents/SIA/

[44]http://www.ivoa.net/documents/SSA/

[45]http://www.ivoa.net/documents/SLAP/

[46]http://www.ivoa.net/Documents/TAP/

[47]http://www.ivoa.net/Documents/latest/ADQL.html

[48]http://www.ivoa.net/documents/ObsCore/

[49]http://www.ivoa.net/documents/SSLDM/

[50]http://www.ivoa.net/documents/Notes/UTypesUsage/index.html

[51]http://www.ivoa.net/documents/ObsCore/

[52]http://www.ivoa.net/documents/ObsCore/