1 Introduction

2 The IVOA Virtual Observatory

The *The International Virtual Observatory Alliance* (*IVOA*) 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 Virtual Observatory (VO) is the realization of the *IVOA* vision of 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 agree the technical standards that are needed to make the VO possible.

The operational VO itself is comprised of a global shared metadata registry, the Registry, and a range of individual data discovery and data access services deployed at each of the participating institutes. These components 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 processes are very similar to the interconnected metadata collections approach described in "The new bioinformatics: integrating ecological data from the gene to the biosphere" (Jones et al. 2006).

> ".... 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 .... 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 data discovery process works in the VO 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 Registry.

The 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 describing all of the services and datasets available 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 for a serviceor dataset has been registered in a local registry, it is automatically propagated up to the next level and replicated between the global registries.

This makes it possible to access the metadata for all of the services and datasets published in the VO by querying any one 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) capability.

In addition to the technical details of services and their capabilities the registry 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, infrared or x-ray.

The registry query would return a table of data, each row of which contains information about a SIA service that provides 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) 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 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 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 location of suitable data followed by individual queries to establish what the data is and how to access it.

3 Tropical forest science

3.1 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) comparing remote sensing data from satellites 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) and the Woods Hole Research Center (Baccini et al., 2012) with four sets of ground plot data from Red Amazónica de Inventarios Forestales (RAINFOR) (Malhi et al. 2002) the Amazon Tree Diversity Network (ATDN) (ter Steege et al., 2003) the Tropical Ecology Assessment and Monitoring (TEAM) network and the Brazilian Program for Biodiversity Research (PPBio).

3.1.1 Satellite data

3.1.2 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) 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), 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.

3.2 Diverse metadata formats

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

3.2.1 Global Index of Vegetation-Plot Databases

The Global Index of Vegetation-Plot Databases (GIVD) 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]` as is the PPBio information system `[GIVD:SA-BR-001]` and the data from the TEAM network `[GIVD:00-00-002]`.

Dengler et al. describe the GIVD project in a 2011 paper, "The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science" (Dengler et al. 2011) 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."

Which is similar to the distributed architectrure of data discovery and data access services used by the virtual observatory.

3.2.2 METACAT

Different institutes have different emphasis and different aproaches to handling the metadata associated with

In a 2012 paper by Flávia Fonseca Pezzini et al. about the PPBio project (Pezzini et al. 2012) "The Brazilian Program for Biodiversity Research (PPBio) Information System" they describe the role of the data manager and the metadata collection processes that are in place.

They also describe the transition from data storage in flat files, which was sufficient for the first five years of the project, to a new system based on Metacat.

> 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 is an open source data management tool that provides a repository for managing both data and metadata in one system.

> 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.

Metacat is capable of handling a variety of different metadata formats, including Ecological Metadata Language (EML) FGDC Biological Data Profile.

3.2.3 DataONE

The Metacat project is itself part of the Data Observation Network for Earth (DataONE) project, a collaboration spconsored by the U.S. National Science Foundation to build an infrastructire from distributed webservices that provides open, persistent, robust, and secure access to Earth observational data.

> 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.

The DataONE arcitecture is based on a set of top level *Coordinating Nodes* and *Member Nodes* located at each participating institute or organisation

Coordinating Nodes provide a replicated catalog of Member Node holdings, enabling scientists to discover data wherever they reside, and data repositories to make their data and services available to the international community.

The individual *Member Nodes* at each institute enable them to make their data available to the rest of the DataONE network via a standard webservice interface.

Again, this two layers of data discovery and data access is similar the virtual observatory architecture.

References

Jones, Matthew B. et al. (2006). "The new bioinformatics: integrating ecological data from the gene to the biosphere". In: *Annual Review of Ecology, Evolution, and Systematics* 37, pp. 519–544. DOI: 10.1146/annurev.ecolsys.37.091305.110031. URL: http://www.annualreviews.org/doi/abs/10.1146/annurev.ecolsys.37.091305.110031.

IVOA. The International Virtual Observatory Alliance. URL: http://www.ivoa.net/.