AstroTROP Evaluation Report

D. Morris
University of Edinburgh

13 May 2015, Version 1.0

1 Introduction

This report draws on the knowledge and experience gained from development of the *AstroGrid* and *IVOA* virtual observatory systems to answer the following related questions:

- What would be required to implement a virtual observatory system for the TROPGLOBE research community, capable of supporting the science use-cases outlined on the *AstroTrop* website.
- What components from the *AstroGrid* and *IVOA* projects would be appropriate to use in developing a virtual observatory system for the TROPGLOBE research community.

2 The IVOA Virtual Observatory

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 number 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 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 range e.g. infrared, visible light, radio or x-ray.

The query may come direct from a user explicitly querying the service, or the may be generated by an application searching for suitable type of data for it to process.

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 registry services, typically hosted at the participating institute level, working in cooperation with a set of higher level global registry services 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 service or 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 *IVOA Simple Image Access* (*SIA*) capability.

In addition to the technical details of services and their capabilities the *Registry* also contains details about the content of datasets, including details of the wavelength(s) measured, e.g. infrared, visible, radio or x-ray.

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 returns 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 list of *IVOA* standards is the *HEALPix Multi-Order Coverage Map* (*MOC*) which allows *Registry* services to perform coarse grained region matches.

This will enable us to further refine our *Registry* 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.

An *SIA* service can handle queries that specifiy a particular wavelength and a particular region of the sky:

- `POS` The positional region (ra, dec).
- `BAND` The energy interval (wavelength).

In the context of this example, a positional point in the sky identified by the right ascension (ra) and declination (dec) polar coordinates are broadly analogous to the terrestrial latitude and longitude coordinate system.

Each *SIA* service returns 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 *Observation Data Model Core Components* (*ObsCore*) data model.

This demonstrates a core part of the *IVOA* architecture, interoperable services based on standard interfaces and data formats.

All of the *SIA* services will return a standard response, which makes it much easier to combine them to produce a global list of all the images available within the whole VO that match our search criteria.

The two key components of this 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 *SIA* 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 is very similar to the stages described in Jones et al. 2006:

1. Querying the metadata to establish the location of suitable data.
2. Querying the individual services 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 *NASA Jet Propulsion Laboratory* (*JPL*) "Benchmark map of forest carbon stocks in tropical regions across three continents" (Saatchi et al. 2011), and the *Woods Hole Research Center* (*WHRC*) "Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps" (Baccini et al. 2012), with four sets of ground plot data from the following sources:

- *Red Amazónica de Inventarios Forestales* (*RAINFOR*) (Peacock et al. 2007) (Malhi et al. 2009).
- *Amazon Tree Diversity Network* (*ATDN*).
- *Tropical Ecology Assessment and Monitoring* (*TEAM*).
- *Brazilian Program for Biodiversity Research* (*PPBio*) (Pezzini et al. 2012).

3.1.1 Remote sensing source data

The paper does not give details of the data discovery and data access methods used to access the primary remote sensing source data. However, there are a number of data discovery tools available that enable researchers to search for remote sensing data products such as satellite images and radar scans.

Good examples of this type of tool are the *Earth Explorer* and *GloVis* tools provided by the *U.S. Geological Survey* (*USGS*).

> "The USGS EarthExplorer ... provides users the ability to query, search, and order satellite images, aerial photographs, and cartographic products from several sources."
>
> "In addition to data from the Landsat missions and a variety of other data providers, EE now provides access to MODIS land data products from the NASA Terra and Aqua missions, and ASTER level-1B data products over the U.S. and Territories from the NASA ASTER mission."
>
> "The USGS Global Visualization Viewer (GloVis) is an online search and order tool for selected satellite data. The viewer allows access to all available browse images from the Landsat 7 ETM+, Landsat 4/5 TM, Landsat 1-5 MSS, EO-1 ALI, EO-1 Hyperion, MRLC, and Tri-Decadal data sets, as well as Aster TIR, Aster VNIR and MODIS browse images from the DAAC inventory."

The *USGS* also provides large area composited mosaics generated from *Landsat* data via the *WELD* project.

> "The WELD data products are processed so users do not need to apply the equations, spectral calibration coefficients, and solar information, needed to convert Landsat digital numbers to reflectance and brightness temperature. They are defined in the same coordinate system and align precisely, making them simple to use for multi-temporal applications. The products provide consistent data that can be used to derive higher-level land cover as well as geo-physical and biophysical products for assessment of surface dynamics and to study Earth system functioning."

The *USGS* also maintains a *Long Term Archive* (*LTA*) of historical remote sensing data:

> "The U.S. Geological Survey's (USGS) Long Term Archive (LTA) at the National Center for Earth Resource Observations and Science (EROS) in Sioux Falls, SD is one of the largest civilian remote sensing data archives."
>
> "Time series images are a valuable resource for scientists, disaster managers, engineers, educators, and the general public. USGS EROS has archived, managed, and preserved land remote sensing data for more than 35 years and is a leader in preserving land remote sensing imagery."

However, all of these interfaces are based around human interaction. There do not apper to be any machine readable data discovery services for this type of remote sensing data.

3.1.2 Carbon density maps

A detailed description of the dataset produced by *NASA Jet Propulsion Laboratory* (*JPL*) is available in the associated paper (Saatchi et al. 2011).

The paper, along with the additional supporting information available on the *Proceedings of the National Academy of Sciences of the United States of America* (*PNAS*) website, provide a textual description of the primary data sources and the analysis methods that were applied.

However, technical details of the data sources, instruments, target areas and date ranges the data covers are not available in a *machine readable* format.

The carbon density dataset itself is available as *GTIF* files, with associated *World file* metadata, for download from the *NASA JPL carbon dataset* site.

A detailed description of the dataset produced by the *Woods Hole Research Center* (*WHRC*) is available in the associated paper (Baccini et al. 2012).

The paper, along with the additional supporting information available from the *Nature* website, provide a textual description of the primary data sources and the analysis methods that were applied.

However, technical details of the data sources, instruments, target areas and date ranges the data covers are not available in a *machine readable* format.

The carbon density dataset itself is available by request from the *WHRC carbon dataset* website. Access to the data requires filling in a simple web form declaring who you are and what you want to use the data for. On submitting the webform, an automated email reply is generated containing a URL to a *ZIP* file on the *WHRC* website.

The *ZIP* file referred to in the email contains the data as *GTIF* files, with associated *World file* metadata.

This is a fairly standard mechanism for providing access to research data. However, it may not be sufficient to support the complex data classification and indexing needed to support an integrated *VO* system.

3.1.3 Ground plot data

The four sets of ground plot data from *RAINFOR*, *ATDN*, *TEAM* and *PPBio* were combined together in the *ForestPlots.Net* database.

Details of the design and capabilities of the *ForestPlots.Net* system is presented in "ForestPlots.net: a web application and research tool to manage and analyse tropical forest plot data" (Lopez-Gonzalez et al. 2011).

> "The ForestPlots.net web application was designed primarily as a repository for long-term intact tropical forest inventory plots, where trees within an area are individually identified, measured and tracked through time."

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 downloaded and imported into the *ForestPlots.Net* database manually.

A permanent archive of the combined field plot data is stored in the *ForestPlots.Net* database as a publically available dataset[1] and is available in the supporting information for the paper.

3.1.4 *AGB* data

The *AGB* data for the forest plots were calculated using a *SQL* query provided by the *ForestPlots.Net* system which implements the tropical forest model described in "Tree allometry and improved estimation of carbon stocks and balance in tropical forests" (Chave et al. 2005). The results of the *AGB* calculation for each forest plot are included in the combined field plot dataset stored in the *ForestPlots.Net* database.

The paper refers to a number of maps derived from the field plot data and other sources which were generated as part of the analysis:

- *Kriged* map of mean wood density (ρ).
- Ratio of diameter (D) to tree height (H) Feldpausch et al. 2012.
- *Kriged* map of basal area.
- *Kriged* map of *AGB* using D and species-specific ρ, and a regional height model ($\mathrm{K_{DH}}\rho$).

[1] `http://dx.doi.org/10.5521/FORESTPLOTS.NET/2014_1`

- *Kriged* map of *AGB* using D and species-specific ρ, but a pan Amazonian height model ($K_{D\rho}$).
- *Kriged* map of *AGB* using D, regional height models and ρ, but with ρ fixed at 0.63 (K_{DH}).
- *Kriged* map of *AGB* using D, pan-Amazonian height model, and ρ fixed at 0.63 (K_{D}).
- *AGB* map from Saatchi et al. 2011 (RS1).
- *AGB* map from Baccini et al. 2012 (RS2).
- Difference between RS1 and $K_{DH\rho}$.
- Difference between RS2 and $K_{DH\rho}$.
- Difference between RS1 and RS2.

These derived datasets and maps are not available in the supporting information for the paper.

The *AGB* data derived from two remote-sensing-derived maps, Saatchi et al. 2011 and Baccini et al. 2012 are not available in the supporting information for the paper.

4 *AstroTrop* requirements

Based on the *AstroTrop* use cases we have studied so far it is clear that data discovery forms a significant part of the requirements for *AstroTrop*.

4.1 External data

In many of the use cases a significant part of the source material for the use case has come from outside the *AstroTrop* community.

For example, both the Saatchi et al. 2011 and Baccini et al. 2012 datasets used in the Mitchard et al. 2014 use case came from external data sources, *NASA Jet Propulsion Laboratory* and *Woods Hole Research Center* respectively.

In the short-term, in order to make this type of external data available as part of the *AstroTrop* data discovery process, it will be necessary for a member of the *AstroTrop* community to register and curate the *AstroTrop* metadata describing the external data.

In the longer-term, the ideal solution would be to encourage external data providers like *NASA Jet Propulsion Laboratory* and *Woods Hole Research Center* to join the *AstroTrop* community and participate in the development of the standards and *web service* interfaces for data sharing and discovery.

It is worth noting that a number of *NASA* projects are active members of the *IVOA*, participating in the working groups and conferences and contributing towards developing the *IVOA* standards.

In order to promote this, it may be beneficial for *AstroTrop* members to establish links with, and become members of, existing international standardization efforts within the relevant communities.

4.2 Internal data

A number of the *AstroTrop* use cases require access to data provided by members of the *AstroTrop* community. Promoting and facilitating this kind of data sharing and re-use of results within the *AstroTrop* community is one of the key goals of the *AstroTrop* project.

In order to support this activity, the *AstroTrop* system needs to enable individual members of the *AstroTrop* community to publish metadata describing their datasets in the *AstroTrop* system.

Once this metadata is available within the *AstroTrop* system, it enables other members of the *AstroTrop* community to discover and use the data as source material for their own research.

5 *IVOA* software

In both cases, the requirements for the data discovery process are that the users are able to specify an area of interest and the type of data they are interested in and then gradually narrow the search criteria in response to the data discovery results until they find the most suitable data for their purposes.

Based on this outline we can begin to evaluate how well the *IVOA* and *AstroGrid* software meets the *AstroTrop* requirements and compare this with equivalent *Geographic Information System* (*GIS*) software available.

At first glance, the *IVOA* and *AstroTrop* data discovery processes are very similar. Suggesting that the *IVOA* and *AstroGrid* software should be a good fit for the *AstroTrop* requirements.

However, there are a number of issues that may mean that the *IVOA* and *AstroGrid* software are not the best solution for meeting the *AstroTrop* requirements.

5.1 Data models

One issue is that a significant part of the *IVOA* metadata structure include a number of domain specific astronomy concepts and terms, making it an imperfect match for a different domain.

Although it would be possible to remove the domain specific concepts and terms from the *IVOA* data model and replace them with something more suited to the *AstroTrop* domain. Doing this piece at a time, gradually evolving a new metadata data model for the *AstroTrop* project would be a non-trivial undertaking involving a significant commitment of time and resources.

It is worth noting that the *IVOA ObsCore* data model that forms the basis of the *IVOA* data discovery process is the result of 10 years' work by the *IVOA* working groups to define a common data model for astronomy observations. It would be likely to take a similar length of time for the *AstroTrop* community to develop an equivalent data model from scratch.

With this in mind, it may be more practical to base the *AstroTrop* metadata on existing data models and data description techniques that are already in use within the *AstroTrop* community or that has been developed for domains closley related to to the *AstroTrop* community.

There are a number of such data models are available. Two examples of these are the *World file GIS metadata* and *Ecological Metadata Language* (*EML*) metadata formats.

5.1.1 *World file GIS metadata*

The *World file GIS metadata* format provides a simple way of annotating an existing map or raster image with *GIS* location metadata.

The *World file* format consists of a plain text file format containing details of the location, scale and rotation of a map or raster image.

Both of the Saatchi et al. 2011 and Baccini et al. 2012 remote sensing datasets provide *World file* metadata using the *example.tfw* convention to associate the metadata with the *GTIF* maps.

This is a simple example of an established convention within the *GIS* community for linking *GIS* metadata to datasets or maps.

5.1.2 *Ecological Metadata Language*

EML is a detailed set of specifications for metadata describing ecological datasets, based on work done by the Ecological Society of America and associated efforts "Nongeospatial metadata for the ecological sciences" (Michener et al. 1997).

These are just two examples of metadata data models and data description techniques that are already in use within the *GIS* and Ecology domains.

This highlights a significant opportunity for the *AstroTrop* community to identify the existing metadata datamodels and data description techniques already in use by members of the *AstroTrop* community and to work together to define a common set of interoperable models and techniques that best describe the data used by the *AstroTrop* community.

5.2 Data owners

A second issue with the *IVOA* and *AstroGrid* metadata *Registry* and data discovery tools concerns the allocation of roles and responsibilities for managing the metadata within the *Registry*, and the way these reflect the structure of the *IVOA* and the members involved in developing the *Registry*.

Historically, the most active contributors to the development of the *IVOA* standards and the *AstroGrid* software have been primary data providers within the international and UK astronomy communities. Many of these represent large scale data providers responsible for publishing and curating primary science archives for telescope surveys or satellite missions.

In the *AstroTrop* domain these are equivalent to the upstream data providers who publish the original satellite remote sensing data, such as the *Landsat* data archive or the *Long Term Archive* (*LTA*) of remote sensing data published by by the *USGS*.

This has influenced the way that the *IVOA* and *AstroGrid* software and services have been developed. In particular, the priority has been to concentrate on providing tools and services for publishing the large primary source datasets.

This emphasis on the larger data providers has meant that the curation of the dataset metadata was seen as a system administrator role. As a result, many of the current tools for managing and curating datasets are designed around a single system administrator role managing the metadata for an entire service, rather that individual researchers managing the metadata for their own data.

In contrast, in the *AstroTrop* use cases the hope is that a significant portion of the data in the system will be provided by and curated by individual researchers or small research groups. For example, the results and supplementary data for the (Mitchard et al. 2014) paper would be published and curated by the members of the research team themselves.

As a result, the structure of the data models and access control systems and the design of the user interfaces of the *IVOA* and *AstroGrid* software and services would need significant work to adapt them to support the new use cases.

6 Alternative software

The design issues identified with the *IVOA* software would not prevent using it as the basis for developing the *AstroTrop* system.

It should be possible to gradually replace the *ObsCore* data model in the *IVOA* software with a new data model designed for *AstroTrop*, and it should be possible to develop a new user interface and permission infrastructure to enable individual users to publish and curate their own data.

On the other hand, there are a significant number of existing software applications and systems which have been specifically designed for handling geographical and ecological data.

Many of these systems may be capable of providing an equivalent level of functionality as the *IVOA* and *AstroGrid* software and it may be useful to look at a few examples to see how they compare.

6.1 Global Index of Vegetation-Plot Databases

The *Global Index of Vegetation-Plot Databases* (*GIVD*) system is a complex registry of metadata describing databases of vegetation plot data from around the world.

The *GIVD* system contains records for over 200 databases and 3 million individual vegetation plots.

Three of the datasets used in our use cases are listed in the *GIVD* system:

- [GIVD:00-00-001][2] *ForestPlots.Net.*
- [GIVD:SA-BR-001][3] *PPBio.*
- [GIVD:00-00-002][4] *TEAM.*

In "The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science" (Dengler et al. 2011) the *GIVD* project team describe the system architecture and some plans for the future to aggregate different types of data from external sources.

> "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."

This is broadly similar to the *VO* architecture of distributed datasets and to the *'interconnected metadata collections approach'* described in "The new bioinformatics: integrating ecological data from the gene to the biosphere" (Jones et al. 2006).

However, the current emphasis is focussed on providing a human interactive search facility, with the *GIVD* system acting as the central node. The current plans do not include providing a machine readable interface to enable the *GIVD* system itself to be used as a component in a larger distributed system.

6.2 *PPBio* Information System

In "The Brazilian Program for Biodiversity Research (PPBio) Information System" (Pezzini et al. 2012) the *PPBio* team describe the role of the data manager and the metadata collection processes developed as part of the *PPBio* Information System.

They also describe the transition from an initial flat file data storage system, 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."

This move towards open standards for both the metadata (*EML*) and the service interfaces (*Metacat*) enables the *PPBio* Information System to become part of a larger distributed system.

6.3 *Knowledge Network for Biocomplexity*

The *Knowledge Network for Biocomplexity* (*KNB*) is a network designed to:

> "facilitate ecological and environmental research"

by enabling researchers to:

> "share, discover, access and interpret complex ecological data."

The *KNB* system is based on a set of *open source* software and standards developed and maintained as part of the *KNB* project:

[2] `http://www.givd.info/ID/00-00-001`
[3] `http://www.givd.info/ID/SA-BR-001`
[4] `http://www.givd.info/ID/00-00-002`

- The *Morpho* data management tools.
- The rDataONE R package for accessing *DataONE* repositories.
- The *Metacat* metadata database.
- The *Ecological Metadata Language* (*EML*) metadata language.

6.3.1 *Metacat*

Metacat is a data management tool that provides a repository for managing data and metadata in a single 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 uses the *EML* metadata data model and vocabulary to describe datasets in the network.

In some cases the *Metacat* system stores both the metadata and actual data itself, e.g. *Tree crown allometries, Piedmont and Southern Appalachians 2001-2004*[5].

In other cases the *Metacat* system only stores the metadata, referring to data that is stored elsewhere, e.g. *Tree crown allometries, Piedmont and Southern Appalachians 2001-2004*[6].

This would be a good match for the *AstroTrop* requirements and use cases.

In some cases, the *AstroTrop* system needs to store both the data and the metadata, e.g the results and supplementary data for (Mitchard et al. 2014).

In this example, a member of the team working on (Mitchard et al. 2014) would store the results and supplementary data in the *AstroTrop* system together with the metadata describing their data.

In other cases, the *AstroTrop* system would only store the metadata, along with a refence to the data stored in an external system, e.g. the (Saatchi et al. 2011) and (Baccini et al. 2012) source material datasets used by (Mitchard et al. 2014).

In this example, a member of the team working on (Mitchard et al. 2014) would enter the metadata details of the (Saatchi et al. 2011) and (Baccini et al. 2012) datasets into the *AstroTrop* system, but the actual datasets would remain with their original publishers, *NASA Jet Propulsion Laboratory* and *Woods Hole Research Center* respectively.

All three datasets would appear to be within the same (virtual) system, enabling *AstroTrop* users to discover the (Saatchi et al. 2011) and (Baccini et al. 2012) remote sensing datasets alongside the results and all of the supplementary data for the (Mitchard et al. 2014) paper.

6.4 *Data Observation Network*

Metacat and the *KNB* project is part of the *Data Observation Network* (*DataONE*) project.

The *DataONE* project is part of the *National Science Foundation* (*NSF*) funded *DataNet* (*DataNet*) programme to build an infrastructire 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."

[5] `https://knb.ecoinformatics.org/#view/doi:10.5063/AA/mdietze.3.2`
[6] `https://knb.ecoinformatics.org/#view/doi:10.5063/AA/mdietze.3.2`

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

The top level *Coordinating Nodes* provide a replicated cataloge of the data in the *Member Nodes*, enabling researchers to search for and discover data across the whole network.

The individual *Member Nodes* at each institute enable researchers to publish data to the whole *DataONE* network.

This hierarchical structure is similar to the *VO* architecture of a global *Registry* containing metadata describing datasets and services in the *VO* and the *'interconnected metadata collections approach'* described in "The new bioinformatics: integrating ecological data from the gene to the biosphere" (Jones et al. 2006).

6.5 *Geospatial Platform*

The *Federal Geographic Data Committee* (*FGDC*) *Geospatial Platform* is designed to :

> "provide a suite of well-managed, highly available, and trusted geospatial data, services, and applications for use by Federal agencies-and their State, local, Tribal, and regional partners."

The *Geospatial Platform* system brings together metadata standards, software and services that provide a set of features which are similar to those that the *AstroTrop* aims to provide.

- Map Viewer.
- Trusted Datasets.
- Multiple Basemaps.
- Collaborative Groups.
- Editable Layers.

6.6 *Comprehensive Knowledge Archive Network*

A key component of the *Geospatial Platform* system is the *Comprehensive Knowledge Archive Network* (*CKAN*) service, which provides the main metadata and data repository for the system.

Development of *CKAN data management system* (DMS) is managed by the *Open Knowledge* network.

CKAN is used to power official data portals by national and local governments in the UK, Brazil, the Netherlands, Austria, the US.

Examples of science based CKAN sites:

- *University of Bristol* Research Data Repository[7].
- *Danube Reference Data and Service Infrastructure*[8].
- *National Energy Technology Laboratory* Energy Data eXchange[9].
- *National Geothermal Data System* [10].
- *National Oceanic and Atmospheric Administration* data catalog[11].

[7] `http://data.bris.ac.uk/data/`
[8] `http://drdsi.jrc.ec.europa.eu/`
[9] `https://edx.netl.doe.gov/about`
[10] `http://geothermaldata.org/`
[11] `https://data.noaa.gov/dataset`

As with the *Metacat* system, *CKAN* is able to store the data along with the metadata describing it, or just store the metadata about a data resource held in an external system.

This matches the two use cases described above, where members of a research team would store their results and supplementary data in the *AstroTrop* system together with the metadata describing the data, or they would enter the metadata for an external dataset stored in remote repository.

As with the *Metacat* system, *CKAN* is designed to function as a node in a federated network of services, using a metadata harvesting mechanism to bring together metadata about resources in other nodes.

This model of using distributed federation of collaborating services is similar to the *IVOA* and *AstroGrid* architecture and the *'interconnected metadata collections approach'* described in "The new bioinformatics: integrating ecological data from the gene to the biosphere" (Jones et al. 2006).

7 Conclusion

7.1 Core software

Although the *AstroGrid* software could in theory be modified to meet the *AstroTrop* requirements, the results of this evaluation indicate that the changes required would be non-trivial.

A number of factors contribute to this, including the differences between the data models, use cases and data ownership, and the fact that the service technologies and standards have evolved and user expectations have changed since the core *AstroGrid* systems were developed.

Conversely, a number of systems already exist in domains adjacent to *AstroTrop* which are based on similar ideas and provide broadly similar functionality.

The *Metacat* and *CKAN* systems are two examples that make use of the latest technical standards and technologies and are perhaps better suited to handling the *AstroTrop* use cases and data models.

With this in mind, we recomend that future work on the *AstroTrop* project looks at basing the *AstroTrop* system on the *Metacat* or *CKAN* systems.

This enables us to build on the considerable geospatial functionality and domain-specific knowledge that is already available within the developer communities for these systems, while at the same time taking full advantage of the key lessons and knowledge gained from our involment in developing the *AstroGrid* and *IVOA VO* projects.

Both *Metacat* and *CKAN* already provide support for the *GIS* and *EML* metadata models. In addition, both projects are actively involved in large scale cross disciplinary data management projects.

- *Metacat* is part of the *DataONE* project, which is specifically aimed at building the infrastructure for earth observational data.
- *CKAN* is used in a number of national *Data.Gov* projects, in particular the *FGDC Geospatial Platform* system which is specifically aimed at handling geospatial data.

A number of the *AstroTrop* science cases are based on primary remote sensing data from sources such as *Landsat* and on processed data products from groups such as *JPL* and *WHRC*. Ideally, in the long term, the best way to support the *AstroTrop* science cases would be to encourage the external data providers to join the *AstroTrop* community and adopt the same metadata and service standards.

Basing the *AstroTrop* system on software that is already being used by cross disciplinary projects that handle earth observational and geospatial data using the established *GIS* and *EML* metadata models is likely to make it easier encourage data providers such as *JPL* and *WHRC* to participate.

The best way to start this process is for the *AstroTrop* community to work together to identify metadata datamodels and data description techniques already in use and to work together to define a

common set of interoperable models and techniques that best describe the data used by the *AstroTrop* community.

In addition, where the *AstroTrop* community relies on existing and established standards, representative members of the *AstroTrop* community should join the relevant external groups responsible for developing these standards.

7.2 Project structure

The *AstroGrid* project succeded in its goal of delivering a *VO* system for the UK astronomy community, and as one of the founding members of the *IVOA* played a significant part in establishing the *IVOA* structure and processes.

The fact that over 10 years after it was first established the the *IVOA* is still an active community of scientists and engineers busy working on developing the next generation of services and specifications is due to the organizational structures and processes put in place early in the *AstroGrid* and *IVOA* projects.

We would recomend that the *AstroTrop* project adopts a similar organizational structure and development processes.

An important lesson learned from the *AstroGrid* and *IVOA* projects is that developing and agreeing common metadata and interoperability standards is crucial to the successful operation of a heterogenous, distributed data system, and that this process must involve all the stake holders working together, including the end user data-consumers and the primary source data-providers.

The key component of this is building an active community and encouraging the participants to work together to design and develop their own set of use cases, standards and data models that meet the community's requirements.

A National *OpenData* projects using CKAN

- Argentina (`http://datospublicos.gob.ar/`)
- Austria (`https://www.data.gv.at/`)
- Australia (`http://data.gov.au/`)
- Canada (`http://open.canada.ca/en`)
- European Union (`http://open-data.europa.eu/en/data/`)
- Germany (`https://www.govdata.de/`)
- Italy (`http://www.dati.gov.it/`)
- Netherlands (`https://data.overheid.nl/`)
- Norway (`http://data.norge.no/`)
- Ireland (`http://data.gov.ie/`)
- Romanina (`http://data.gov.ro/`)
- Slovakia (`http://data.gov.sk/`)
- Switzerland (`http://opendata.admin.ch/`)
- Uganda (`http://www.data.ug/`)
- UK (`http://data.gov.uk/`)
- USA (`http://www.data.gov/`)

References

AGB. Above Ground Biomass. URL: http://www.ipcc-nggip.iges.or.jp/public/2006gl/pdf/4_Volume4/V4_04_Ch4_Forest_Land.pdf.

ATDN. Amazon Tree Diversity Network. URL: http://web.science.uu.nl/Amazon/ATDN/.

AstroGrid. URL: http://www.astrogrid.org/.

AstroTrop. URL: http://www.astrotrop.org/.

Baccini, A. et al. (2012). "Estimated carbon dioxide emissions from tropical deforestation improved by carbon-density maps". In: *Nature Climate Change* 2 (3), 182–185. DOI: 10.1038/nclimate1354. URL: http://www.nature.com/nclimate/journal/v2/n3/full/nclimate1354.html.

PPBio. Brazilian Program for Biodiversity Research. URL: http://www.teamnetwork.org/.

Chave, J. et al. (2005). "Tree allometry and improved estimation of carbon stocks and balance in tropical forests". In: *Oecologia* 145 (1), pp. 87–99. DOI: 10.1007/s00442-005-0100-x. URL: http://link.springer.com/article/10.1007/s00442-005-0100-x.

CKAN. Comprehensive Knowledge Archive Network. URL: http://ckan.org/.

DRDSI. Danube Reference Data and Service Infrastructure. URL: http://drdsi.jrc.ec.europa.eu/.

DataONE. Data Observation Network. URL: https://www.dataone.org/.

Data.Gov. URL: http://www.data.gov/.

DataNet. URL: https://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503141.

Dengler, Jürgen et al. (2011). "The Global Index of Vegetation-Plot Databases (GIVD): a new resource for vegetation science". In: *Journal of Vegetation Science* 22 (4), pp. 582–597. DOI: 10.1111/j.1654-1103.2011.01265.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2011.01265.x/abstract.

Dietze, Michael, Michael Wolosin, and James Clark (2004). *Tree crown allometries, Piedmont and Southern Appalachians 2001-2004.* URL: https://knb.ecoinformatics.org/#view/doi:10.5063/AA/mdietze.3.2.

Earth Explorer. URL: https://lta.cr.usgs.gov/earth_explorer.

EML. Ecological Metadata Language. URL: https://en.wikipedia.org/wiki/Ecological_Metadata_Language.

FGDC. Federal Geographic Data Committee. URL: https://www.fgdc.gov/.

Feldpausch, T. R. et al. (2012). "Tree height integrated into pantropical forest biomass estimates". In: *Biogeosciences* 9 (8), 3381–3403. DOI: 10.5194/bg-9-3381-2012. URL: http://www.biogeosciences.net/9/3381/2012/bg-9-3381-2012.html.

ForestPlots.Net. URL: http://www.forestplots.net/.

Kriged. Gaussian process regression. URL: https://en.wikipedia.org/wiki/Kriging.

GIS. Geographic Information System. URL: http://en.wikipedia.org/wiki/Geographic_information_system.

Geospatial Platform. URL: http://www.geoplatform.gov/.

GTIF. GeoTIFF. URL: http://trac.osgeo.org/geotiff/.

GIVD. Global Index of Vegetation-Plot Databases. URL: http://www.givd.info/.

GloVis. Global Visualization Viewer. URL: https://lta.cr.usgs.gov/glovis.

MOC. HEALPix Multi-Order Coverage Map. URL: http://www.ivoa.net/documents/MOC/.

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

Registry. IVOA Registry. URL: http://www.ivoa.net/documents/RegistryInterface/.

SIA. IVOA Simple Image Access. URL: http://www.ivoa.net/documents/SIA/.

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.

KNB. Knowledge Network for Biocomplexity. URL: https://knb.ecoinformatics.org/.

Landsat. Landsat Program. URL: http://landsat.gsfc.nasa.gov/.

LTA. Long Term Archive. URL: https://lta.cr.usgs.gov/about.

Lopez-Gonzalez, Gabriela et al. (2011). "ForestPlots.net: a web application and research tool to manage and analyse tropical forest plot data". In: *Journal of Vegetation Science* 22 (4), pp. 610–613. DOI: 10.1111/j.1654-1103.2011.01312.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2011.01312.x/abstract.

machine readable.

Malhi, Y. et al. (2009). "An international network to monitor the structure, composition and dynamics of Amazonian forests (RAINFOR)". In: *Journal of Vegetation Science* 13 (3), pp. 439–450. DOI: 10.1111/j.1654-1103.2002.tb02068.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2002.tb02068.x/abstract.

Metacat. URL: https://knb.ecoinformatics.org/knb/docs/intro.html.

Michener, William K. et al. (1997). "Nongeospatial metadata for the ecological sciences". In: *Ecological Applications* 7 (1), pp. 330–342. DOI: 10.1890/1051-0761(1997)007[0330:NMFTES]2.0.CO;2. URL: http://www.esajournals.org/doi/abs/10.1890/1051-0761%281997%29007%5B0330%3ANMFTES%5D2.0.CO%3B2.

Mitchard, Edward T. A. et al. (2014). "Markedly divergent estimates of Amazon forest carbon density from ground plots and satellites". In: *Global Ecology and Biogeography* 23 (8), pp. 935–946. DOI: 10.1111/geb.12168. URL: http://onlinelibrary.wiley.com/doi/10.1111/geb.12168/abstract.

Morpho. URL: https://www.dataone.org/software-tools/morpho.

JPL. NASA Jet Propulsion Laboratory. URL: http://carbon.jpl.nasa.gov/.

NASA JPL carbon dataset. URL: ftp://www-radar.jpl.nasa.gov/projects/carbon/datasets/.

NASA. National Aeronautics and Space Administration. URL: http://www.nasa.gov/.

NETL. National Energy Technology Laboratory. URL: http://www.netl.doe.gov/.

NGDS. National Geothermal Data System. URL: http://geothermaldata.org/.

NOAA. National Oceanic and Atmospheric Administration. URL: http://www.noaa.gov/.

NSF. National Science Foundation. URL: http://www.nsf.gov/.

Nature. URL: http://www.nature.com/.

ObsCore. Observation Data Model Core Components. URL: http://www.ivoa.net/documents/ObsCore/.

OKFN. Open Knowledge. URL: https://okfn.org/.

open source. URL: https://en.wikipedia.org/wiki/Open_source.

OpenData. URL: https://en.wikipedia.org/wiki/Open_data.

Peacock, J et al. (2007). "The RAINFOR database: monitoring forest biomass and dynamics". In: *Journal of Vegetation Science* 18 (4), 535–542. DOI: 10.1111/j.1654-1103.2007.tb02568.x. URL: http://onlinelibrary.wiley.com/doi/10.1111/j.1654-1103.2007.tb02568.x/abstract.

Pezzini, Flávia Fonseca et al. (2012). "The Brazilian Program for Biodiversity Research (PPBio) Information System". In: *Biodiversity & Ecology* 4 (24), 265–274. DOI: 10.7809/b-e.00083. URL: http://www.biodiversity-plants.de/biodivers_ecol/article_meta.php?DOI=10.7809/b-e.00083.

PNAS. Proceedings of the National Academy of Sciences of the United States of America. URL: http://www.pnas.org/.

RAINFOR. Red Amazónica de Inventarios Forestales. URL: http://www.rainfor.org/.

Saatchi, Sassan S. et al. (2011). "Benchmark map of forest carbon stocks in tropical regions across three continents". In: *Proceedings of the National Academy of Sciences of the United States of America* 108 (24), 9899–9904. DOI: 10.1073/pnas.1019576108. URL: http://www.pnas.org/content/108/24/9899.

SQL. Structured Query Language. URL: https://en.wikipedia.org/wiki/SQL.

TEAM. Tropical Ecology Assessment and Monitoring. URL: http://www.teamnetwork.org/.

University of Bristol. URL: http://www.bris.ac.uk/.

USGS. U.S. Geological Survey. URL: http://www.usgs.gov/.

VO. Virtual Observatory. URL: http://www.ivoa.net/.

WELD. Web-enabled Landsat data. URL: http://landsat.usgs.gov/WELD.php.

web service. URL: https://en.wikipedia.org/wiki/Web_service.

WHRC carbon dataset. URL: http://www.whrc.org/mapping/pantropical/carbon_dataset.html.

WHRC. Woods Hole Research Center. URL: http://www.whrc.org/.

World file. World file GIS metadata. URL: https://en.wikipedia.org/wiki/World_file.

ZIP. ZIP archive. URL: https://en.wikipedia.org/wiki/Zip_(file_format).