ASTROTROP: ACHIEVEMENTS AND NEXT STEPS

A Discussion Paper

Alan Grainger
University of Leeds

1. Introduction

ASTROTROP is a unique collaboration between astronomers and global change scientists. Its primary aim has been to evaluate the potential to adapt AstroGrid/Euro-VO virtual observatory software so that members of the TROPGLOBE network can use it to overlay tropical forest datasets in a virtual Pan-Tropical Forest Observatory.

The ASTROTROP project is funded by the Science and Technology Facilities Council's Challenge-Led Applied Systems Programme (CLASP), which aims to: apply STFC-Funded Research to have impacts in STFC Global Challenge Areas, which include the environment; fill technology gaps, which include automated monitoring of ecosystems and carbon emissions; produce "demonstrable deliverables for a potential market", including scientists, companies, policy makers and other stakeholders; and have environmental, social and commercial benefits.

This discussion paper is being circulated as a basis for discussion at "Measuring the Planet 2016", the Second ASTROTROP Conference. It reports the findings of the evaluation exercise, its recommendations for initial demonstrations of the virtual observatory concept, and suggestions for the practical next steps which TROPGLOBE members can take to realize the goal of a virtual Pan-Tropical Forest Observatory, by agreeing on standards and principles of collaboration.

This document is based on the work of Dave Morris, Keith Noddle and Andy Lawrence at the Royal Observatory, Edinburgh, which they reported at a workshop in Leeds in July 2015, and on consultations which Alan Grainger has since undertaken with TROPGLOBE members and other interested parties.

The key conclusion of the AstroGrid/Euro-VO Evaluation Report is that this software cannot be used directly, or easily adapted for a new use. Instead, it was recommended that tropical forest researchers imitate the AstroGrid approach by: (a) using existing open geospatial standards devised by the Open Geospatial Consortium (OGC), and (b) agreeing on their own standards to make existing OGC compliant software functional for their needs.

The good news is that because a huge amount of work has already been put into assembling standards for open geospatial data, and a range of software compliant with these standards is already available, the standards on which TROPGLOBE members need to agree on will fall largely in the scientific domain, rather than in the technical (IT) domain. It is recommended that TROPGLOBE follow the example of the International Virtual Observatory Alliance (IVOA), and set up Working Groups for standards in key areas. Breakout groups at the Conference will make a start in agreeing on which standards are needed, and which Working Groups should be established to specify these standards formally. IVOA and OGC standards are summarized in two appendices of this paper for information only. The Conference will focus on scientific standards, not IT standards.

Another topic for discussion at the conference is on other future actions to continue the work undertaken in ASTROTROP. One of these actions will involve using the ASTROTROP approach in existing and future funded projects. Another action will be to seek funding to widen our networking activities: the change in focus of this initiative, from a software process to a standards process, means that we can shorten the time needed to agree on standards by networking with other initiatives. This will include intensifying our networking activities with commercial, policy and other stakeholders, to see how our new approach can benefit them.

If we are to improve our understanding of the role of tropical forests in global environmental change it is vital that we start to observe Planet Earth just like we observe other planets. As the ASTROTROP project has showed, there is much that we can learn from astronomers in improving Earth Observation, and our conference will also seek new ideas in this area too.

Another aim of the ASTROTROP project was to build a community of tropical forest scientists out of the many disciplines that study tropical forests. The new approach outlined here, following the results of the AstroGrid/Euro-VO Evaluation Report, shows that community building is as important as it ever was. Global change science will flourish if its members have easy access to global empirical scientific data and information. The virtual Pan-Tropical Forest Observatory will provide an essential tool to facilitate exchanges of data and information within the TROPGLOBE community. It will not result in single estimates of tropical forest attributes, but enable different groups to develop and test new methods to refine these estimates. Astronomers have shown us how to use community facilities to make great strides in science. It is now up to us to realize this potential in tropical forest research.

2. Background

Since remote sensing satellites have been collecting global data since 1972, ideally these data would have been converted into global digital information on the distribution of forests, other ecosystems and land use which scientists could analyse. Instead, we lack regular measurements of changes in world forest area, and global environmental change scientists have had to rely on compilations of unreliable national statistics. This has limited the accuracy of scientific research into global environmental change, and of information available to policy makers.

The idea of a World Forest Observatory was devised to overcome this information deficiency by bringing together scientists from different disciplines, who would collaborate in producing annual maps of world forest area, biodiversity and carbon density which they, and other scientists worldwide, could analyse. The initial plan for a World Forest Observatory in 2010 relied on a *hub design*, in which 10-20 teams across the world would produce digital maps of three forest attributes; area, biodiversity and carbon. These maps would then be combined on a central geographical information system at a global hub, from which users would download global digital maps. However, in an alternative *virtual design*, the digital maps of multiple forest attributes would remain on the information bases of the teams which produce them, but sophisticated software would enable the multiple attribute maps to be overlaid in geo-referenced format in the computers of users which download them. In addition, the relatively small number of core teams in the original approach would be replaced by a much larger cooperative community which studies a wider range of forest attributes, including ecosystem processes, gaseous processes and hydrology. The importance of a community approach is clearly demonstrated by the growing use of remote sensing data to estimate carbon stocks in forests, thereby eroding the previous compartmentalization between remote sensing for forest area monitoring and ground studies for forest carbon monitoring. Here, as in other areas, multiple communities will benefit from exchanging data, e.g. for calibration and validation.

Establishing a Pan-Tropical Forest Observatory in the UK as a pilot for a World Forest Observatory is highly appropriate since UK scientists are leaders in tropical forest research. It would also act as a focus for a vibrant industry devoted to developing the innovative technologies needed for a growing family of global environmental observatories that can ensure a seamless chain from satellite data to information on policymakers' desks, consistent with the goals of the UK Satellite Applications Catapult, and of the European Commission's Copernicus Programme.

3. How the AstroGrid/Euro-VO Virtual Observatory Works

A virtual observatory can be summed up as having "all the world's databases inside your PC", just as the Worldwide Web means that "all the world's documents are inside your PC". Users can search for data and services, and then combine and overlay data, images, catalogues, spectra and time series. Discovering the data that you need is facilitated by the metadata attached to each parcel of data, which describe, for example, wavelength, location and pixel size.

AstroGrid virtual observatory software has been refined through European partnerships under the umbrella of the International Virtual Observatory Alliance of astronomers (IVOA). It includes processes, documents, data, services and code:

i. *Processes* include twice yearly international meetings of the standards body, and working groups and interest groups which agree on standards documents. Processes are modelled on the Worldwide Web Consortium (W3C), and involve both the IVOA and the multidisciplinary Research Data Alliance (RDA). Standards agreed by the standards body are normally ratified by the IVOA generally.

ii. *Documents* include: data models which define how data are structured, and standards for expressing location in space and time, registry metadata, service protocols and data query language.

iii. *Data* include: registries (or Yellow Pages), collections of images and spectra, table-like databases, and events.

iv. *Services* include registry querying, image and spectrum grabbing, database querying, and event querying. After data sources are identified in registries other services are used to grab data from identified locations.

v. *Code* includes user tools, such as Topcat and Aladin; code for running services; and libraries.

Consequently, the astronomical virtual observatory is not mainly about software, but the result of a huge community effort by dozens of people over 13 years to agree on the standards needed to make it possible. Some of the most important IVOA Standards are summarized in Appendix 1. Standards are important for applications, services and data exchange. If tropical forest scientists wish to replicate AstroGrid/ Euro-VO then they need to replicate the communal efforts of astronomers.

The standards are produced by Working Groups and Interest Groups. The IVOA has Working Groups on Applications; Semantics; Working Groups: Data Access Layer; VO Event; Data Modelling; Resource Registry; Grid and Web Services; and VOTable. It has Interest Groups on: Theory; Education; Operations; Data Curation and Preservation; and Knowledge Discovery in Databases. Other Groups and Committees include the Standing Committee on Standards & Processes and the Standing Committee on Science Priorities (http://wiki.ivoa.net/twiki/bin/view/IVOA).

Virtual observatory standards evolved gradually, with simple things being tackled first, and members of the TROPGLOBE network could emulate this approach. This would enable data and information sharing in the network to become operational as quickly as possible, and make the Pan-Tropical Forest Observatory operational too.

4. Main Findings of the AstroGrid/Euro-VO Evaluation Report

The needs of global change scientists for spatial information on tropical forests are similar to those of astronomers, which is why the case for refining AstroGrid/Euro-VO software for use in a Pan-Tropical Forest Observatory originally seemed so compelling. However, the main conclusions of the evaluation are that:

i. Whereas there are a lot of large astronomical datasets, which can be accessed through a virtual observatory because their hosts are members of the IVOA, the same is not true for global change science. The data in most of the large datasets, such as those of NASA, NOAA, ESA etc., are not directly usable by global change scientists without intermediate processing into geospatial information. For the most part, geospatial information needed by global change scientists is held by individual researchers and relatively small research groups and networks.

Table 1. Geospatial equivalents of the Components of AstroGrid/Euro-VO software.

		Astronomy	Geospatial Equivalents
i	*Data models*	Obscore	Worldfile GIS metadata Ecological Metadata language
ii.	*Standards bodies*	IVOA	Open Geospatial Consortium
iii.	*Registries*	AstroGrid/Euro-VO Registry	Comprehensive Knowledge Archive Network (CKAN)
iv.	*Registry and Data Grabbing Services*	SIAP	Geoserver
v.	*Data Manipulation Tools*	Aladin	QGIS

ii. Information required by TROPGLOBE partners falls into two main categories: (a) information within the TROPGLOBE network, which can be obtained by controlled access, while testing the use of metadata and publishing services; and (b) information outside the TROPGLOBE network, which can be obtained by establishing links with the hosts of the requisite databases.

iii. AstroGrid/Euro-VO software and standards are not reusable without significant modification and time because they are too specific to the needs of astronomers.

iv. The structure of the AstroGrid/Euro-VO project is, however, reusable if TROPGLOBE partners have the collective will to replicate the efforts made by astronomers in making a virtual observatory feasible, e.g. by establishing processes to agree on standards and describe these in documents, making their public datasets consistent with these standards, and writing code for data/information sharing.

iv. Nevertheless, because other members of the geospatial and environmental science communities have been pursuing the open data concept for some time now, sufficient open source geospatial standards have been agreed by the Open Geospatial Consortium (OGC), and sufficient software has been devised using the standards, to form the basis for prototyping a Pan-Tropical Forest Observatory. Table 1 shows the geospatial equivalents of the components of AstroGrid/Euro-VO software.

CKAN is used by Open Data Germany (http//: open-data.fokus.fraunhofer.de/en/ernten-und-geerntet-werden-erfahrungen-beim-govdatadeharvesting/) and by the US National Geothermal Data System (www.geothermal-energy.org/pdf/IGAstandard/SGW/2013/Clark.pdf). Other key geospatial initiatives include: (i) Global Index of Vegetation-Plot Databases (http://www.givd.info), a registry containing metadata for 226 vegetation plot databases and links to them; (ii) Data Observation Network for Earth (DataONE), a registry that aims to store, and give access to multi-scale, multi-discipline, and multi-national science data (https://www.dataone.org/software-tools/tags/GIS); and (iii) Knowledge Network for Biocomplexity, a network of federated institutions within DataONE which share data and metadata using a common framework, and Ecological Metadata Language as a

common language to describe ecological data (http://www.dcc.ac.uk /resources/implementations/knb-knowledge-network biocomplexity). Other global data initiatives with which it will be valuable to network include, but are not restricted to: (i) ICSU World Data System; (ii) WMO Global Atmosphere Network; (iii) Taxonomic Data Working Group (TDWG); (iv) World Resources Institute Global Forest Watch; and (v) Global Biodiversity Information Facility (GBIF). National bodies that have also made important strides in agreeing on vegetation standards include the US Federal Geographic Data Committee.

5. A Blueprint for a Prototype Pan-Tropical Forest Observatory

This section outlines a blueprint agreed at the Leeds Workshop in July 2015 for a prototype Pan-Tropical Forest Observatory.

5.1 Registry Software

The first component is the Registry, which links Databases and connects them, in turn, to Users.

Three candidates for Registry Software are: CKAN, GeoNodes and Metacat. CKAN has been chosen because it is available now; does what the IVOA Registry does; has an adaptable metadata store; is widely used in open data services in the USA; and is being continually updated by a large active community. The University of Bristol, an ASTROTROP partner, uses a CKAN service to structure its data repository: data.bris.ac.uk. Each partner could have its own repository which forms a sub-directory that is part of the central directory of an overall repository. Each dataset should be described by a metadata file that includes such elements as: (i) attribute; (ii) date created; (iii) date last updated; (iv) data type; (v) spatial coordinates.

5.2 Registry and Data/Information Grabbing Services

Once the locations of required data or information are identified they can be extracted from their host databases and channelled to a user's computer. The chosen candidate for Data/Information Grabbing Software is Geoserver. This can "cut out" from a large geospatial database the data/information for an area defined by latitude and longitude coordinates, and is consistent with GeoTIFF and CSW formats. CKAN itself does not handle GIS data but has a plug-in extension for Geoserver. Geoserver is consistent with six standards for data formats, so it should be suitable. Partners, and end users, might expect to see outputs in Computable Document Format (CDF) and in GeoTIFF format, which embeds georeferencing information.

5.3 User Tools

Once data have been channelled to a user's computer they can be analysed using standard GIS software tools, such as ArcINFO, QGIS and R.

5.4 Alternative Architectures

Other architectures are possible too. For example, a few years ago Duncan Golicher of the University of Bournemouth proposed using GeoServer to channel open

information layers to the computers of end-users, where they would be stored on PostGIS information bases and analysed using R software.

5.5 Demonstrating the New ASTROTROP Approach

TROPGLOBE members will be able to use a facility provided on the ASTROTROP website to test the use of this approach for overlaying their data.

It would be help if TROPGLOBE members use a standard template to describe their overlays. This could include:

i. Primary researcher
ii. Science goal
iii. Datasets
iv. Problem description
v. Current solution
vi. VO solution
vii. Spatial cover
viii. Spatial extent
ix. Formats for metadata, input and output information files
x. Validation and errors protocols
xi. Policy relevance

6. Next Steps: Topics for Discussion

Activities since the AstroGrid/Euro-VO Evaluation Report was released show the exciting potential for realizing a virtual Pan-Tropical Forest Observatory, and reveal some important topics for discussion at our conference on the best ways to do this.

6.1 Standards and Working Groups

One of the main items on the agenda of the Second ASTROTROP Conference will be to agree on initial temporary standards and on an optimal set of Working Groups to generate permanent standards for exchanging and overlaying data.

The vast majority of these standards will be scientific in content since OGC software standards have been evolving for some time. Special file types for inputs and outputs are not needed because open data standards recognize them all. Scientists have also become accustomed to adopting open data standards in their regular practices by giving URL and DOI numbers to datasets and databases.

The key types of standards which seem to be needed, and possible names for the associated Working Groups, include:

i. Semantics, to agree on terms, definitions and existing vocabularies for forest area, biodiversity, biomass, carbon, ecology, gases, hydrology etc.

Key terms include: (a) forest, tree, shrub, bush; (b) forest area, tree cover, tree density; (c) aboveground biomass, carbon density, burnt area.

ii. Metadata, to identify primary forest attributes, features tailored to each attribute, and quality and other criteria which are crucial for data searches, so that common metadata files can be attached to data files.

iii. Registries, to promote cooperation between the multiple existing registries which already exist.

iv. Query terms, to identify key query approaches which different users will want to use, and determine if they can be encompassed in existing generic query protocols.

v. Applications, to agree on standards in other key areas, including table formats etc.

All of these activities will recognize that there is no one way to define any term or search for any data, but that different approaches should be identified so that there can be translation between them.

Detailed examples of sets of metadata are provided in Appendices 1 and 2. But one set of general metadata could include, but would not be restricted to:

a. Source.
b. Abstract.
c. Title.
d. Spatial resolution.
e. Frequency of upload.
f. Geographical projection.

A more specialized set of metadata for carbon estimates could include, but would not be restricted to:

a. Plot coordinates acquired with GPS.
b. Based on ground measurements after the year 2000.
c. AGB in living trees above a certain threshold dbh.
d. Allometric model used.

Various vocabularies, or glossaries, already exist, and the most important ones should be identified so that they can be utilized.

6.2 Practicalities of Implementation

Inspection of existing database systems and open data implementation reveals a number of issues regarding the practicalities of implementing a virtual observatory.

Some scientific teams disseminate their data through their own organization's website, but others do so through larger 'nodes', or portals, such as the World Resources Institute's Global Forest Watch, which 'harvest' data from individual websites, and have greater accessibility or user-friendly features. Such nodes are entirely consistent with a virtual observatory approach, but good communication between all groups involved is still essential.

Different approaches are also taken to providing the software needed for virtual observatories. One approach is to distribute software over multiple servers but another approach is to place all software on the same server.

Registries, or "Yellow Pages", are a fundamental part of any virtual observatory. But as with any website it will be necessary to have some form of administration, and two questions are where should that administration be based and how should it operate?

In contrast to astronomy, tropical forests are studied by many disciplines, so agreeing on standards will be more complicated because standards must be acceptable to the different needs of members of all disciplines that engage in interdisciplinary tropical forest research. In some areas, specialist Working Groups may be needed for each community, together with an integration working group to link things together.

Experience has shown that implementing a virtual observatory, or any open data system, works best if there is wide consultation beforehand and participation in the actual implementation. This will help to recognize the needs of different groups and actors and ensure that all needs and potential problems are identified.

This shows that virtual observatories, and open data systems generally, are human systems and not merely IT systems. Just as personal practicalities must be accommodated, so too must the needs of the intellectual approaches of all the disciplines involved. Structuring data and information should be seen as not just an arbitrary IT role, but as a scientific role too.

6.3 Concerns

Consultations leading up to the conference have revealed a number of concerns by tropical forest researchers about the practicalities of open data. These concerns, which will be discussed at the conference, include:

i. Openness is often a one-way street. Users want data from one member of a network, but are not prepared to share their own data in return.

Hopefully, the TROPGLOBE community can agree on principles of collaboration, for example, reciprocal sharing, acknowledging the source of data (including DOI) in citations, having sources as co-authors if they wish etc.

There could be differences in sharing responsibilities: forest area data, for example, could be the most open, because it provides the base map for all other attributes, while data on other attributes could be less open and more regulated.

ii. Data suppliers receive insufficient feedback on how often their data are used, especially on nodes.

One way to address this is for nodes to have automatic systems for despatching metrics to data sources, and for users to be encouraged to contact original data sources if they have any queries.

iii. It costs money to curate and service databases.

This issue needs to be addressed at various levels, including ASTROTROP but also research councils. ASTROTROP can tackle this by seeking support for proposals to fund database curation.

iv. The long-term availability of data is not guaranteed.

This issue could be tackled if ASTROTROP provided some kind of automatic backup system.

v. Links to datasets can get broken or become confusing, for example, as URLs change, datasets are refined, or species names change through taxonomic refinements. It can be difficult to keep track of the version numbers corresponding to different datasets.

Part of this problem can be addressed by how files are structured and related. Ultimately, however it is a curation issue for which proper administration and funding should be sought.

6.4 Future Projects

Since STFC funding for this phase of ASTROTROP's activities is coming to an end it will be important to discuss at the conference the possibilities for continuing its activities in existing projects and for applying for funding for new projects.

Kevin Tansey, of the University of Leicester, will give an overview at the conference of how he proposes to use the ASTROTROP approach is part of the GlobBiomass project.

Our original ASTROTROP proposal mentioned that we would include in our final report a proposal and budget for a Pan-Tropical Forest Observatory, and we shall use the discussions at the conference as a basis for doing this.

Other possibilities include applying for funding for: (a) joint projects involving the funding of both ASTROTROP and AstroGrid/Euro-VO; and (b) EU Horizon 2020 European Research Infrastructure projects. Suggestions for other funding opportunities will be welcomed at the conference.

There is huge potential to use the ASTROTROP approach in the commercial arena as well as for a wide range of other stakeholders. Because the results of the AstroGrid/Euro-VO Evaluation Report showed that existing technology could not be easily adapted for ASTROTROP needs it has not been possible to engage with stakeholders within this project as much as was originally intended. However, the new approach that has been devised following the publication of the Evaluation Report offers even wider possibilities than were originally anticipated, and exploring these could also lead to funding for future activities.