Executive Summary on the 2018 Astrolabe Datathon

March 1-2, 2018
Biosphere 2, Oracle, Arizona
https://osf.io/buer6/

Workshop sponsored by:

National Science Foundation

Workshop organizing committee:

Bryan Heidorn, University of Arizona School of Information
Gretchen Stahlman, University of Arizona School of Information
Julie Steffen, American Astronomical Society

Astrolabe Co-PI's

Bryan Heidorn, University of Arizona School of Information
Julie Steffen, American Astronomical Society

Workshop participants:

Amanda Bauer	Large Synoptic Survey Telescope
Eiichi Egami	University of Arizona Department of Astronomy and Steward Observatory
Erik Enderlein	Usermind
Katy Garmany	National Optical Astronomy Observatory
John Glaspey	National Optical Astronomy Observatory
Bryan Heidorn	University of Arizona School of Information
Tom Hicks	Tohono Consulting, LLC / University of Arizona
Chris Kollen	University of Arizona Libraries
Jim Martin	University of Arizona Libraries
Susan Miller	CyVerse
Fernando Rios	University of Arizona Libraries
Phil Rosenfield	Harvard-Smithsonian Center for Astrophysics
Greg Schwarz	American Astronomical Society
Gretchen Stahlman	University of Arizona School of Information
Julie Steffen	American Astronomical Society
Frank Timmes	Arizona State University / American Astronomical Society
Ramona Walls	CyVerse

Introduction and Purpose

As documented in Heidorn, Stahlman & Steffen (2018), the Astrolabe project began as a collaboration between University of Arizona and the American Astronomical Society to create a resource for curation of astronomical data not otherwise curated elsewhere. While the discipline of astronomy has pioneered best practices in management of large and diverse datasets (Henneken, 2015; Accomazzi & Dave, 2011; Hanisch, et al., 2007), many researchers possess so-called "dark", at-risk or potentially at-risk data, and these researchers could benefit from new curation resources (Heidorn, 2008). A series of community workshops (see Heidorn, Stahlman & Steffen, 2018; Stahlman, Heidorn & Steffen, 2018) informed creation of Astrolabe as a project within CyVerse cyberinfrastructure [1], tasked with developing tools and interfaces to make the CyVerse system useful for astronomers as a convenient place to deposit, analyze and visualize astronomical data. CyVerse features include proximity to HPC, sophisticated data-sharing architecture, and virtual environments for software implementation - all of which encourage data discovery, reproducibility and re-use, overall in support of open science initiatives. Furthermore, the system can accommodate old and new data, while allowing processing and reduction of raw data in addition to intermediate data products.

Workshops held in 2015 and 2016 led to recommendations from the astronomical community for Astrolabe system development, and resulted in a funded National Science Foundation proposal. A key recommendation of these workshops was to leverage existing cyberinfrastructure, and particularly CyVerse as a National Science Foundation investment. The Astrolabe Project was subsequently awarded a three-year NSF Software Infrastructure for Sustained Innovation (SI2-SSE) grant[2], beginning in 2016 and with objectives to:

- Develop WorldWide Telescope (WWT) software as a visual front-end for the Astrolabe repository of legacy data;
- Create tools for processing raw data into Astrolabe and visualization formats needed for WWT visualization;
- Integrate the Unified Astronomy Thesaurus (UAT) taxonomy into both Astrolabe and WWT;
- Implement data manipulation tools necessary to use WWT as a first look in archive browsing and retrieval.

CyVerse cyberinfrastructure can support these objectives by providing longevity through a supported platform for data storage, as well as curating software within virtual environments, minting permanent identifiers (DOIs) for datasets to be linked to publications, and permitting robust metadata.

Now in Year 2 of the project, Astrolabe is engaging users and obtaining feedback on design. The purpose of the 2018 Datathon described here was primarily to unveil and test the following tools that have been developed for Astrolabe thus far:

[1] cyverse.org

[2] National Science Foundation Award #1642446: Visualizing Astronomy Repository Data using WorldWide Telescope Software Systems; P.B. Heidorn, PI, J. Steffen, Co-PI

- **JS9**: An online version of DS9[3], integrated into the Astrolabe website[4]
- **FITS Extractor**: An application in the CyVerse Discovery Environment; automatically extracts FITS headers into a metadata template
- **FITS Validator**: A web tool that validates FITS headers and assists users in correcting errors through the Astrolabe website
- **Metadata Templates**: Three metadata templates are now available for Astrolabe users - a template for FITS data following the IVOA ObsCore standard[5]; a template for minting DOIs following the DataCite standard[6]; and a Dublin Core template[7]
- **Unified Astronomy Thesaurus (UAT)**[8]: UAT terms are suggested when a user beings typing a keyword in a CyVerse metadata template
- **Atmosphere**: An Astrolabe Virtual Machine instance is pre-configured with astronomical data analysis software software and easy connection to the CyVerse Data Store
- **WWT on the Web**: The WWT web client has been optimized for web use and repository linking through the WWT API

Datathon Structure

The 2018 Astrolabe Datathon was held at University of Arizona's Biosphere 2 near Tucson, AZ. The location was chosen for its comfortable conference facilities, as well as the unique character of this isolated venue[9]. Designed to be an immersive experience, participants spent two days and nights at the Biosphere 2 for a face-to-face workshop and hackathon, considering and providing feedback on Astrolabe system development and existing and future tools.

Invitations to the Datathon were emailed to 55 individuals who were identified by the workshop organizing committee as experts in relevant fields and institutions. A variety of local and out-of-town invitees were included, primarily comprised of astronomers, but also including librarians and computational scientists. The NSF grant specified that the workshop would include 24 participants (14 out of town and 10 local); however, the project needed to stay within budget considering the Biosphere 2 cost estimate. Furthermore, we unknowingly scheduled the Datathon at the same time as Steward Observatory's prospective graduate student visits, so participation was therefore limited.

The final Datathon team included 17 participants (mostly local to the Tucson area, and four out-of-town guests). The moderate size of the group was ideal and permitted meaningful discussion and hands-on work. A mix of expertise was represented, including:

- 7 astronomers
- 3 librarians
- 2 cyberinfrastructure specialists

[3] http://ds9.si.edu/site/Home.html
[4] astrolabe.arizona.edu
[5] http://www.ivoa.net/documents/ObsCore/
[6] https://schema.datacite.org/
[7] http://dublincore.org/
[8] http://astrothesaurus.org/
[9] http://biosphere2.org/

- 2 programmers
- 2 information science researchers
- 1 scholarly publisher

The agenda for the two-day Datathon is included in Appendix A. Collaborative notes and documentation from the workshop are collected in the Open Science Framework.[10]

Outcomes and Recommendations

The community consistently affirms a need for Astrolabe, but the project must address the following issues:

- The distinction between Astrolabe, WWT and CyVerse is not clear:
    - "Easy to get lost"
    - Not obvious which tools we are using in CyVerse and which are independent of CyVerse
- Tools need more work, new functionality
- CyVerse has issues also:
    - Search doesn't work
    - Atmosphere takes a long time to start up, loses connection
    - The generic CyVerse interfaces are frustrating for astronomers, and Astrolabe should develop customized interfaces for astronomy data
- Need to engage users through targeted efforts (see below), by creating collaborative projects with astronomers, and hosting training "schools"
- Need both user documentation and code documentation (Open Science Framework can be a platform for connecting project management, team members and stakeholders, publications, with documentation and code)

Objectives Moving Forward

- Fix tools:
    - FITS header extractor - needs to extract all metadata, not just WWT fields
    - Need script to link Atmosphere instance to data store without substantial user efforts
    - Create interface and index for Astrolabe data through the astrolabe.arizona.edu web portal to compensate for CyVerse search issues:
        - Indexing now is several steps through the FITS header extractor to create a .csv file then index; this should be transparent to users
        - Perhaps API to link to an index on the Astrolabe web page
- Develop "quick-start"/decision-tree guides based on use cases:
    - Uploading published data sets ready for DataCite and DOI - Needs both single file and multi-file uploads
    - Upload unpublished FITS files data that may need metadata validation to view in WWT. Do not overwrite but add to metadata - VO Index

[10] https://osf.io/sp349/

    - Manipulate data with cloud tools such as Astropy, Jupyter Notebooks, etc. running on our Astrolabe Atmosphere Instance
    - Upload images with no FITS header - use astrometry.net and custom edit tools to create fits header
    - Browse, search and view data already in the repository and view in WWT
- WWT as discovery for Astrolabe:
    - With API, visual discovery tool for data on website
    - VO registry implementation - can search Astrolabe data in WWT
    - Create Astrolabe community in WWT
- Make additional software available in CyVerse:
    - Astroconda, Jupyter Notebook - this is now in Astrolabe Atmosphere instance.
    - Needs use cases and testing.
- Target user communities:
    - Deep field/JWST community
    - Serious amateur astronomers
    - Smaller telescopes such as university facilities that might need web services for data
    - Create and host a school or training program (similar to MESA Summer School model[11]) to engage graduate students, postdoctoral researchers, faculty, amateurs
- Find funding to support additional development beyond the scope of SI2 grant, which is WWT-centric
- Engage librarians as proxy to handle metadata and assist with data upload for researchers at their institutions
    - Need CyVerse/Astrolabe training

Appendix: Astrolabe Datathon Agenda

Wednesday February 28

7:00 Dinner (Cafe)

Thursday March 1

7:30-9:00 Breakfast, Cafe
9:00 Workshop begins, Sahara Room
9:00-9:30 Introduction - Astrolabe (Bryan Heidorn, Gretchen Stahlman - UA School of Information)
9:30-9:50 CyVerse (Susan Miller, CyVerse)
9:50-10:00 Discussion
10:00-10:20 Astrolabe Tools (Erik Enderlein, former Astrolabe Software Developer)
10:20-10:30 Discussion
10:30-10:40 Break

[11] http://cococubed.asu.edu/mesa_summer_school_2018/

10:40-11:10	WorldWide Telescope (Phil Rosenfield or Julie Steffen, American Astronomical Society)
11:10-11:20	Discussion
11:20-11:40	CyVerse Discovery Environment data upload
11:40-12:00	Discussion and morning summary
12:00-1:00	Lunch, Sahara Room
1:00-1:20	Meta-data (Ramona Walls, CyVerse)
1:20-1:40	Discussion and data annotation
1:40-2:00	Atmosphere (Bryan Heidorn)
1:50-2:20	Discussion and cloud data processing hands-on
2:20-2:40	Open Science Framework (Fernando Rios, UA Library)
2:40-4:00	Hands on linking data in OSF
4:00-6:00	Biosphere Tour (Tour guide will meet us in the Sahara Room and then drop us off at the Dome for dinner after the tour)
6:00	Dinner in the Biosphere Lower Habitat

Friday March 2

7:30-9:00	Breakfast, Cafe
10:00	Workshop reconvenes, Sahara Room
10:00 - 10:15	AAS update (Julie Steffen)
10:15 - 11:30	Breakout groups on Astrolabe design
11:30-12:00	Discussion
12:00-1:00	Lunch, Sahara Room
1:00-3:00	Hunting Party for Dark Data (Astronomers mainly, but everyone can observe and participate in whole-group discussion)
3:00	Workshop ends

References

Accomazzi, A. and Dave, R. (2011). Semantic interlinking of resources in the Virtual Observatory era. Astronomical Data Analysis Software and Systems XX. *ASP Conference Proceedings, 442*, 415.

Hanisch, R.J., Steffen, J., Choudhury, S., Dilauro, T., Szalay, A., Vishniac, E., Ehling, T., Milkey, R. & Plante, R. (2007). Digital data preservation and curation: A collaboration among libraries, publishers, and the Virtual Observatory. *Library and Information Services in Astronomy V: Common Challenges, Uncommon Solutions. ASP Conference Series*, 377.

Heidorn, P.B., Stahlman, G.S., & Steffen, J. (2018). Astrolabe: Curating, Linking and Computing Astronomy's Dark Data. *Astrophysical Journal Supplement Series.* (In Press) https://arxiv.org/abs/1802.03629

Heidorn, P.B. (2008). Shedding light on the dark data in the long tail of science. *Library Trends 57*(2), 280-299.

Henneken, E. (2015). Unlocking and sharing data in astronomy. *Bulletin of the Association for Information Science and Technology, 41*(4), 40-43.

Stahlman, G.S., Heidorn, P.B., & Steffen, J. (In Press). The Astrolabe Project: Identifying and curating astronomical ‘dark data’ through development of cyberinfrastructure resources. *Proceedings of Library and Information Services in Astronomy (LISA) VIII*, Strasbourg, France, June 6-9, 2017.