Executive Summary on the Arizona Astronomical Data Hub Workshop

July 6-7, 2015
University of Arizona, Tucson, Arizona

Report prepared by:

P. Bryan Heidorn, PhD; heidorn@email.arizona.edu
Gretchen Stahlman; gstahlman@email.arizona.edu

Workshop sponsored by:

University of Arizona Office for Research & Discovery
American Astronomical Society

Workshop organizing committee:

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

Workshop participants:

Alberto Accomazzi	Tim Axelrod	Michael Bruck	Eric Christensen
Jeremy DeBarry	Chris Erdmann	Mike Fitzpatrick	Jeremy Frumkin
John Glaspey	Richard Green	Robert Hanisch	Bryan Heidorn
Sharon Hunt	Chris Kollen	Kaitlin Kratter	Derek Masseth
Tom Matheson	Robert McMillan	Susan Miller	Dmitry Mishin
Clayton Morrison	Gus Muench	Kim Patten	Greg Schwarz
David Silva	Edwin Skidmore	Rob Sparks	Gretchen Stahlman
Julie Steffen	Betty Stobie	Kelland Thomas	

Introduction

Astronomy data represent a curation challenge for information managers, as well as for astronomers. Extracting knowledge from these heterogeneous and complex datasets is particularly complicated and requires both interdisciplinary and domain expertise to accomplish true curation, with an overall goal of facilitating reproducible science through discoverability and persistence. A group of researchers and professional staff at the University of Arizona held several meetings during the spring of 2015 about astronomy data and the role of the university in curation of that data. The group decided that it was critical to obtain a broader consensus on the needs of the community. With assistance from a Start for Success grant provided by the University of Arizona Office of Research and Discovery and funding from the American Astronomical Society (AAS), a workshop was held in early July 2015, with 28 participants plus 4 organizers in attendance.

Representing University researchers as well as astronomical facilities and a scholarly society, the group verified that indeed there is a problem with the long-term curation of some astronomical data not associated with major facilities, and that a repository or "data hub" with the correct functionality could facilitate research and the preservation and use of astronomy data. The workshop members also identified a set of next steps, including the identification of possible data and metadata to be included in the Hub. The participants further helped to identify additional information that must be gathered before construction of the AADH could begin, including identifying significant datasets that do not currently have sufficient preservation and dissemination infrastructure, as well as some data associated with journal publications and the broader context of the data beyond that directly published in the journals. Workshop participants recommended that a set of grant proposal should be developed that ensures community buy-in and participation. The project should be developed in an agile, incremental manner that will allow consistent community growth from the early stages of the project, building on existing iPlant infrastructure (www.iplantcollaborative.org) initially developed for the biology community.

Problem Definition: Dark Data, Gray Data, Bright Data

There is an unacceptable quantity of astronomy data that once generated do not have a repository where they can be readily deposited, accessed, reused and preserved. Data from large, planned observation from earth-based and space-based observatories are typically placed in data repositories associated with the telescope and space missions. Very large volumes of such data are stored in those repositories. However, derived data products from the main observations and data from observations from non-mission aligned and unfunded instruments often do not meet the criteria for deposit in those repositories, including some data behind published articles in journals of the American Astronomical Society and other publishers. Data that is published along with journal articles is readily accessible and can be call "bright data" since it is very visible to the scholarly community. Even this data can be improved upon through the additional of additional context that would improve both the verifiability and

replicability of the results. This context includes for example a more detailed specification of the larger dataset from which the published data was derived as well as the methods that may have been used to select, process and display that data. Gray data is not linked to publications but should be available from the original data producers. The data is usually recently produced. It is either provided on a lab website or is available by request from the author. This astronomical data are unique but not associated with traditional publications. These data are maintained by the researchers and amateur astronomers responsible for collecting the data. Such home-curated data are not visible to other researchers around the world and can be considered "dark data". These data are much more likely to be eventually lost for lack of efficient preservation methods and eventual bit-rot.

Workshop structure

A central objective of the Arizona Astronomical Data Hub project is to broaden access to "bright data" and to bring "gray data" and "dark data" into an environment where it will be properly curated and readily accessible to the astronomical community with standard metadata and in standard data formats, including outreach and training directed towards repository users and University of Arizona students. Based on dialogue with members of the astronomy indicating that generic repositories are insufficient for capturing all relevant data while mission archives exist for mission data, the idea of creating a new archive for "orphan" astronomy data was suggested in a series of meetings at University of Arizona and involving American Astronomical Society between 2013 and 2015. Furthermore, a search of the Web of Science database seems to indicate that a significant portion of astronomical research is conducted by scientists affiliated with Arizona institutions. The AADH workshop explored possibilities for creating a new data hub at the University of Arizona.

Workshop invitees included astronomy and data science experts, primarily affiliated with University of Arizona, NOAO and American Astronomical Society, with several invited speakers, and remote participation by individuals at the Center for Astrophysics (CfA) at Harvard University. The appendices at the end of this document contain a full list of participants, as well as the workshop agenda, invitation letter, and informed consent form that all attendees were required to read and sign. Prior to the workshop, invitees were asked to complete an informal survey inquiring about issues with curation of astronomy data. Eleven responses provided initial topics for discussion that were incorporated into the structure of the workshop (see Appendix E). The workshop itself loosely followed the Delphi Method of achieving group consensus (Keeney et al., 2001). Thirty-two participants (remote and in-person) and facilitators were organized into groups based on expertise, ensuring that each group contained a representative mix of astronomers, data scientists, librarians, and educators/administrators. Following an initial introduction by guest speakers, four breakout sessions were held over two days, each immediately followed by whole-group report-out sessions. Detailed notes were taken by each facilitator and by participants using the collaborative note-taking platform Hackpad[1], and invited

[1] https://hackpad.com

talks and whole-group discussions were recorded; this material was analyzed and coded to identify the following broad themes and detailed objectives for the AADH project.

Workshop outcomes and recommendations

1. *Identify mission and clear science use cases*

One or more detailed science use cases are essential to obtaining support and creating a viable Data Hub. Therefore, establishing and documenting potential science cases is a key objective of this grant proposal. AADH workshop discussions indicate that certain existing survey datasets are in danger of becoming “dark data” and could illustrate the utility of a new astronomy data repository based at University of Arizona. In addition, we will mine data from the literature as discussed above.

2. *Take advantage of iPlant infrastructure and longevity of University of Arizona*

iPlant representatives have agreed to dedicate cyberinfrastructure resources to management of the survey data. Leveraging iPlant to support these data and associated software will help us identify and incorporate necessary metadata and data for relevant astronomical datasets. It will also be possible to study participation in this network, as well as broader applications for the astronomy and astrophysics communities. This pilot project will prove that the Arizona Astronomical Data Hub can host 3D image data and handle sophisticated data cubes as well as provide access to requested portions of data, overall earning science advocacy. The Arizona Astronomical Data Hub will take advantage of both the iPlant infrastructure as a computing environment, as well as the astronomy activities on campus and the longevity and stability of the University of Arizona as an institution, to establish a secure and robust repository that focuses on the needs of its users, overall supporting user control of assets and open access to research data, with a commitment to sustainability and education. However, as indicated by Sands, et al. (2014), disciplinary expertise is essential to support curation efforts. As a result, the interdisciplinary Arizona Astronomical Data Hub project team will include a postdoctoral research astronomer who is deeply familiar with techniques for astronomy data mining and visualization as well as existing resources typically utilized by astronomers. Metrics will be established throughout the project to measure success.

3. *Obtain community buy-in and manage expectations*

To solicit buy-in and to ensure that the systems being built meet community needs, we have established an **Advisory Board** for the project. The board will meet at least quarterly during the project. Some board members will have very active roles helping to incorporate their own data sets and advise project staff. The AADH would be part of a complex data space including many ongoing and completed observation missions as well as data service projects and virtual observatories. Lessons learned from these projects were communicated at the AADH workshop and in publications (Hanisch, et al., 2015). As astronomers agreed in the workshop in June, we are attempting to perform cultural engineering, changing the publishing habits of scientists to meet new open access requirements. Of prime importance is to coordinate with the resources at the Harvard-Smithsonian Center for Astrophysics (CfA), and several AADH project members

have already established collaborative relationships with CfA staff. Chief resources to coordinate with include: CfA SAO/NASA Astrophysics Data System (ADS) and Dataverse. Furthermore, the Unified Astronomy Thesaurus (UAT) was recently updated with help from CfA and can be used in AAHD, providing an industrial-strength taxonomy for semantic enrichment. In order to connect to existing resources such as CFA and CDS we have included travel funds for conference where we can meet with representatives of several projects, and we plan to host two additional workshops, with the goal of educating the broader community about the new iPlant resources as well as obtaining feedback from the community of users about their desires for the project's next steps through extramurally funded development. Finally, education is a critical element for the viability and long-term sustainability of the AADH. This includes two main populations: astronomers who deposit data, and students in astronomy and information science who need to learn the techniques. Both the Astronomer Postdoc and the library staff will work with authors of the data to ensure that data are in standard format and with appropriate metadata. PI, Bryan Heidorn and a School of Information doctoral student will work to adapt existing documents to develop best practices guidelines for astronomy data.

4. *Focus on low-hanging fruit*

Workshop participants identified two obvious pieces of "low-hanging fruit" as a niche opportunity for the AADH to provide valuable services to the astronomical community: "dark data" and other orphan datasets not curated elsewhere, and data associated with Arizona authors of articles published in AAS journals. As noted by Henneken and Accomazzi (2012), publications based on data sets are merely expressions of data. Journal publishers are innovating new methods of digital publication that provide rich scientific data beneath a text publication itself. Furthermore, as citation rates appear to be higher for publications that contain links to referenced data, participating in data citation and persistent linking is an important objective of the AADH (Accomazzi, 2011; Accomazzi & Dave, 2011). Additional synergistic activities include working with Microsoft-funded WorldWide Telescope (WWT), as AAS is now in a planning process to make WWT a community-based tool for research, publishing, education and public outreach. iPlant would be an outstanding partner, and the iPlant software stacks are very similar to others in use across the physical sciences community. Additionally, AAS is working with Sloan Foundation to develop a community-based software discovery portal for astronomy with robust developer workflows, unique identifiers, software citation, search and developer credit. We will pilot an instance of this discovery portal in the data repository.

5. *Develop a follow-on workshop*

Workshop participants advised a holding a subsequent workshop to connect community members as the systems develop. This workshop or series of workshops should include both an education component and continued assessment of community needs. These workshops should include data carpentry plus hackathon, targeted faculty with relevant datasets.

Objectives for the future

The Arizona Astronomical Data Hub intends to develop a one-year demonstration project using iPlant cyberinfrastructure resources, and perhaps incorporating Catalina Sky Survey data as a test dataset, overall focusing on interdisciplinary collaboration, coordinating with existing projects, and promoting synergy within the astronomy and scholarly communication ecosystems. Astronomers widely acknowledge data curation issues as a community problem requiring improved communication and standardization to effectively manage increasingly large and complex datasets at various stages throughout the research process. A universal solution does not currently exist for efficiently connecting research astronomers with relevant data, nor have astronomers adopted a single paradigm for curation. Virtual Observatory tools and standards have provided a first step towards widespread best practices (Hanisch, et al., 2015), but additional effort is needed to create adequate repositories and workflows. Renowned computer scientist Jim Gray's "Laws of Data Engineering" reflect the rapid computational changes that affect the process of research and discovery in astronomy (Hey, Tansley & Tolle, 2009), implying that tools must be created and adopted quickly in order to effectively address broad issues with data management. The competitiveness of this proposed project lies in the expertise and funds assembled, the powerful existing capabilities of iPlant, the enthusiastic pledged support of American Astronomical Society, and a concise timeline with clear deliverables to produce and assess a pilot solution for the benefit of the astronomical community overall. Furthermore, iPlant is already engaged with the School of Information through the Applied Cyberinfrastructure Concepts course, which is incorporating astronomy data for student projects and could be further leveraged in the 2016-2017 academic year.

References

Accomazzi, A., & Dave, R. (2011). Semantic Interlinking of Resources in the Virtual Observatory Era, *442*, 10. Retrieved from http://arxiv.org/abs/1103.5958

Accomazzi, A. (2011). Linking Literature and Data: Status Report and Future Efforts, (2010), 9. doi:10.1007/978-1-4419-8369-5_15 http://arxiv.org/abs/1103.4295

Goodman A, Pepe A, Blocker AW, Borgman CL, Cranmer K, Crosas M, et al. (2014) Ten Simple Rules for the Care and Feeding of Scientific Data. *PLoS Computational Biology, 10* (4): e1003542. doi:10.1371/journal.pcbi.1003542

Henneken, E. & Accomazzi, A. (2012). Linking to data: Effect on citation rates in astronomy. *ASP Conference Series*, 461, 763-766. Retrieved from http://arxiv.org/abs/1111.3618

Hey, Tony, Stewart Tansley, & Kristin Tolle (Eds.). (2009). The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research. Retrieved from http://research.microsoft.com/en-us/collaboration/fourthparadigm/default.aspx

Keeney, S., Hasson, F., & McKenna, H. P. (2001). A critical review of the Delphi technique as a research methodology for nursing. International Journal of Nursing Studies, 38(2), 195–200. doi:10.1016/S0020-7489(00)00044-4.

Pepe, A., Goodman, A., Muench, A., Crosas, M. & Erdmann, C. (2014). How do astronomers share data? Reliability and persistence of datasets linked in AAS publications and a qualitative study of data practices among US astronomers. *PLOS One, 9* (8), 1-11.

Sands, A. E., Borgman, C. L., Traweek, S., & Wynholds, L. A. (2014). We're Working On It: Transferring the Sloan Digital Sky Survey from Laboratory to Library. *International Journal of Digital Curation, 9*(2), 98–110. doi:10.2218/ijdc.v9i2.336.

Wynholds, L., Fearon, J., David, Borgman, C., & Traweek, S. (2011). When use cases are not useful: Data practices, astronomy, and digital libraries. Paper presented at the *11th annual international ACM/IEEE joint conference on digital libraries*, 383-386. doi:10.1145/1998076.1998146

Appendix A: Workshop Participants

Name	Organization
Accomazzi, Alberto	NASA Astrophysics Data System
Axelrod, Tim	Steward Observatory
Bruck, Michael	University Information Technology Services
Christensen, Eric	Lunar and Planetary Laboratory
DeBarry, Jeremy	iPlant
Dey, Arjun	National Optical Astronomy Observatory
Erdmann, Chris	Center for Astrophysics
Fitzpatrick, Mike	National Optical Astronomy Observatory
Frumkin, Jeremy	University of Arizona Libraries
Glaspey, John	National Optical Astronomy Observatory
Green, Richard	Steward Observatory
Hanisch, Robert	National Institute of Standards and Technology
Heidorn, Bryan*	UA School of Information
Hunt, Sharon	National Optical Astronomy Observatory
Kollen, Chris*	University of Arizona Libraries
Kratter, Kaitlin	Steward Observatory
Masseth, Derek	University Information Technology Services
Matheson, Tom	National Optical Astronomy Observatory
McMillan, Robert	Lunar and Planetary Laboratory
Miller, Susan	Biocomputing Facility
Mishin, Dmitry	San Diego Supercomputer Center
Morrison, Clayton	UA School of Information
Muench, Gus	American Astronomical Society
Patten, Kim	Office for Research and Discovery
Schwarz, Greg	American Astronomical Society
Silva, David	National Optical Astronomy Observatory
Skidmore, Edwin	iPlant
Sparks, Rob	National Optical Astronomy Observatory

Stahlman, Gretchen*	UA School of Information
Steffen, Julie*	American Astronomical Society
Stobie, Betty	National Optical Astronomy Observatory
Thomas, Kelland	UA School of Information

* *Workshop organizers and breakout group facilitators*

Appendix B: Workshop Agenda

Arizona Astronomy Data Hub Workshop Agenda

July 6 - 9 am to 6 pm

- 9:00-9:30: Catered breakfast in meeting room
- 9:30- 9:45: Welcome and announcements (Julie Steffen)
- 9:45-10:00: Statement from NOAO Director (David Silva)
- 10:00-10:15: Participant introductions
- 10:15-10:30: Overview of objectives for workshop (Bryan Heidorn)
- 10:30-10:45: Life Sciences perspective (iPlant)
- 10:45-11:15: Guest Speaker (Bob Hanisch, NIST)
- 11:15-12:00: Group discussion, introduction to afternoon breakout sessions
- 12:00-1:30: Catered lunch in meeting room
- 1:30-2:30: Breakout #1 (Exploring science needs)
- 2:30-3:00: Report out
- 3:00-3:30: Coffee break
- 3:30-4:30: Breakout #2 (Challenges and opportunities)
- 4:30-5:00: Report out
- 5:00-5:45: Breakout #3 (Exploring current data practices)
- 5:45-6:00: Report out and instructions for Day 2
- 6:00: Day 1 concludes

July 7 - 9 am to 12 noon

- 9:00-9:30: Catered breakfast in meeting room
- 9:30-10:30: Breakout #4 (Prioritizing activities)
- 10:30-11:30: Group discussion (Focus on the future)
- 11:30-12:00: Workshop conclusion

Appendix C: Invitation Letter to Participants

Subject: Workshop for the Arizona Astronomy Data Hub Design, Tucson July 6-7

Invitation letter for Participants.

Dear [Insert Name],

We are pleased to invite you to participate in an important workshop to be held **July 6-7, 2015** at the University of Arizona Library, Room A313/A314. You are invited because your of your experience with astronomy data practices and related data management projects. We are bringing together approximately 35 experts to develop buy-in, key principles and a white paper to use in extramural funding proposals leading to development of an international astronomy data hub for previously uncurated data. There is an unacceptable quantity of astronomy data that once generated do not have a repository where they can be readily deposited, accessed, reused and preserved. Data from large, planned earth-based and space-based observatories are typically placed in data repositories associated with the telescope and space missions. However, derived data products from the main observations and data from observations from non-mission aligned and unfunded instruments often do not meet the criteria for deposit in those repositories, including some data behind articles published in journals of the American Astronomical Society (AAS) and other publishers. Other data are unique but not associated with traditional publications. These data are maintained by the researchers and amateur astronomers responsible for collecting the data. Such home-curated data are not visible to other researchers around the world and can be considered “dark data”; often more likely to be lost for lack of efficient preservation methods and eventual bit-rot. This data are unique and useful to other researchers.

We propose to design and build a “trusted repository” – a repository capable of reliably storing, migrating, and providing access– for these data at the University of Arizona (UA). In considering possible challenges and opportunities for this repository, external experts will travel to University of Arizona (UA) for coordination with UA researchers in order to develop a set of proposals to fund this hub for astronomy data. UA is a world leader in astronomical research. We can build on that reputation and expertise by constructing a self-supporting, long-lived data repository for astronomical data. American Astronomical Society has already agreed to fund a doctoral student for the summer to plan and run this

workshop and help develop the subsequent grant proposals. We plan to leverage the cyberinfrastructure investments at UA in the **iPlant Collaborative, the UA Library** and the **University Information Technology Service** to economically support large datasets and associated computational assets. A demonstration project is already underway utilizing the **Catalina Sky Survey** (http://www.lpl.arizona.edu/css/) inventory of near-earth objects (NEOs) and iPlant CI. With the new UA School of Information, an iSchool, we can leverage researchers and students who are involved in education, research and development to address the data deluge and the associated computational and social resources to advance the field.

Please respond to Bryan Heidorn (**si-research@email.arizona.edu**) and **let us know by June 19, 2015** whether you can attend. If you cannot attend please feel free to return a brief biographical sketch and contact information for someone else from your organization that might bring a similar perspective and knowledge on data management and communication practices in astronomy. Logistics information and a detailed agenda will be sent once we have confirmed the participant list.

The workshop, sponsored by the Office of Research and Development at UA and American Astronomical Society, will explore the software and cyberinfrastructure needs of astronomers, and address how a new repository might advance these objectives. Workshop leaders include Bryan Heidorn from the University of Arizona School of Information, Julie Steffen from American Astronomical Society, Phil Pinto from the University of Arizona Department of Astronomy and Steward Observatory, Chris Kollen from University of Arizona Libraries, and Jeremy Frumkin from University of Arizona Libraries.

Combined, the individuals invited to participate in this workshop represent a breadth of domain and disciplinary expertise. Thus, we anticipate that the workshop deliberations will be engaging, fun, and productive as we share and synthesize diverse perspectives. We appreciate your consideration of our invitation and very much hope that you will be able to join us for this exciting workshop. If you have any questions, please feel free to speak with or email any of us.

Sincerely,

Bryan Heidorn, U. of Arizona School of Information
Julie Steffen, American Astronomical Society
Phil Pinto, Department of Astronomy and Steward Observatory

Chris Kollen, University of Arizona Libraries
Jeremy Frumkin, University of Arizona Libraries

Appendix D: Informed Consent Form

Informed Consent for Arizona Astronomy Data Hub Workshop

Purpose

We will hold a workshop to develop white papers leading to grant proposals for construction of a national hub for astronomical data, a resource currently unavailable. We may publish the findings of the workshop so that the astronomy community will be aware of the opinions of the scientists that attended and the decisions that were made. The hub would provide a data repository at University of Arizona (UA) researchers and researchers from around the globe. The structure, function and content of this repository will be decided in part by the outcomes of this workshop. For administrative work such as this, human subjects informed consent is not needed, but since we intend to publish parts of this workshop informed consent is required by the University of Arizona.

What will happen during this study?

In the week prior to the workshop you will be sent a worksheet with a set of open-ended questions. Please return the worksheet at least two days before the workshop. The organizers will anonymize, collate and organize all of the ideas from the participants to produce summary documents and follow-up questions for the workshop. The 1.5-day meeting will be structured around group discussions and breakout sessions examining specific aspects of the data management issues faced by the astronomical research community. All participants will sometimes meet as one working group and at other times will break into 3-5 working groups. Each session will be motivated by a set of questions addressed first with a brainstorming writing exercise in a focus group format. Participants will be asked to elect one or more "reporters" for each group. This reporter will record individual contributions in a Wiki provided for each group, although anyone may participate in the writing. When breakout group sessions end, a group spokesperson will report out to the entire group. This will be followed by exercises to prioritize outcomes.

Are there any risks to me?

The activities that you will participate in have no criminal, social, financial, or breach of confidentiality risk. Your name will be used in publications or documents only if you leave your name in the notes for the session.

Will there be any costs to me?

Aside from your time, there are no costs for taking part in the study. We will provide catered food during the workshop.

Will I be paid to participate in the study?
You will not be paid to participate in this study. Out of town visitors will be reimbursed for expenses.

Will video or audio recordings be made of me during the study?
Audio recordings will be made during the workshops and sometimes transcribed, but you may request not to be recorded at any time. The audio recordings will not be kept as part of the record but transcriptions may be kept for later analysis. All data produced by this project will be retained both online (wiki and online database) and offline (e.g. DVD copies) for the duration of the conceptualization project and a minimum of 12 months following the project's completion. Public data will be deposited in the University of Arizona Digital Repository, likely in dLIST. If the Arizona Astronomy Data Hub is funded, management of the conceptualization project data will be transferred to the AADH for its duration. If the project is not funded, each partner institution is free to retain or discard the project data after the 12-month period at the local PI's discretion.
The project wiki, which contains all textual data and all descriptions of text and database products, will contain sections available only to project members and as well as publicly accessible regions during the project and for 12 months following the project.

We expect that the project's data will be used by the project team to produce a proposal for the AADH, by sponsors to evaluate our project and the institute proposal, and by AADH stakeholders to validate and otherwise critique our results.

What if I am harmed by the study procedures?
There is no reason to expect any physical harm from the study procedures. There is no reason to expect any psychological harm either, but if you find some of the questions stressful or upsetting, you can stop participating immediately.

May I change my mind about participating?
Your participation in this study is voluntary. You may decide to not begin or to stop the study at any time. All notes will be kept as cloud documents. You may choose to delete or modify any of the notes pertaining to your participation. Also any new information discovered about the research will be provided to you. This information could affect your willingness to continue your participation.

Whom can I contact for additional information?
You can obtain further information about the research or voice concerns or complaints about the research by calling the Principal Investigator **Bryan Heidorn** at (520) 621-3536. If you have questions concerning your rights as a research participant, have general questions, concerns or complaints or would like to give input about the research and can't reach the research team, or want to talk to

someone other than the research team, you may call the University of Arizona Human Subjects Protection Program office at (520) 626-6721. If you would like to contact the Human Subjects Protection Program by email, please use the following email address http://orcr.arizona.edu/hspp.

Signing the consent form

I have read (or someone has read to me) this form, and I am aware that I am being asked to participate in a research study. I have had the opportunity to ask questions and have had them answered to my satisfaction. I voluntarily agree to participate in this study.

I am not giving up any legal rights by signing this form. I will be given a copy of this form**.**

Printed name of subject	**Signature of subject**
	AM/PM

Appendix E: Informal Survey Questions and Results

Initial Report
Last Modified: 07/05/2015

1. 1. Some astronomy data are not currently curated in long-lived databases - in your opinion, what are some of the key technical and social issues associated with this "dark data"?

Text Response
Inadequate human resources to process and document it. Loss of institutional memory of data properties. Lack of academic or professional credit for work on dark data that won't yield peer-reviewed publications. Unreadable legacy media. Scattered locations of the data. Loss of software needed to make sense of old data.
Coming up with a minimum set of meta-data for each deposit so that it can be effectively found via virtual observatory tools. Then you have to help authors reach those minimum and deposit the data. There are also very complex data sets that are difficult to host, think mini-SDSS type databases that include spectra, images, figures and tables all integrated by a search engine. How do you integrate something like that?
o People don't know where to put their data or how to properly tag it with metadata. o We lack a common infrastructure for people to utilize for long-term storage. o Some people remain unwilling to share data.
The key issue is cost/benefit. Not all unpublished ground-based data should be curated indefinitely. Unlike space data with uniform conditions and calibrations, extensive accompanying comment is required to determine the degree of atmospheric transparency and uniformity. Spectroscopic data are subject to variable slit losses from seeing and differential refraction. The social issue is scientific opportunity cost. Reobservation of a very faint target can be expensive in telescope time. Missing a point in a time domain sequence can impact the interpretation, or require negotiation with a person that the investigator did not necessarily intend to be a co-author. Another social issue is whether a completely privately funded unpublished observation obligates the investigator to release that data to some publicly accessible archive, and if so, on what timescale. As a point of discussion, the highest priority could be assigned to the data underlying refereed publications, along with medium-size datasets with uniform calibration that do not currently go to specific ground-based archives.
Technical: - final data products are not properly documented with complete metadata - data formats are not archive/database friendly, e.g. a "working"

table format that doesn't map well to a searchable database - processing history of data is unknown Social: - astronomers don't want to properly separate final data products from intermediate files in their "working directory". A final measurement may come from a file having some unique pattern in the name, but is mixed in with all the files leading to the creation of that spectrum and only the astronomer understands the naming scheme. - cleaning out the working directory then leads to a problem of documenting the data files or putting them in archival form, often seen as an unnecessary extra step since it is the published table of measurements, not the images/spectra from which they were measured, which is important
Technical Issues: --Shall the data be downloadable, or should it be analyzed in situ? --The data will be in a variety of different data formats, each potentially requiring additional software to manipulate and/or extract relevant quantities --The data could potentially take up an enormous amount of space --How to make such a heterogeneous set of data searchable? extractable? Social: --How much of the responsibility of extracting relevant data shall be put on the submitting researcher? --Is there expectation to keep the datasets up to date after they have been initially submitted?
finding the data identifying authors/owners of the data making authors/owners of the data aware of long term archive and share options long term sustainability of archive and share data management compatibility and readability of data over time with changing format/encoding standards provenance of data
- conversion to digital format for truly old data - access (even digital data may be on "legacy" formats like 6250 tape) - utility (not all of it is useful - who decides?) Philosophical issue: shouldn't all federally funded data be instantaneously public?

Statistic	Value
Total Responses	8

2. 2. What existing astronomy data repositories do you use? Please mention some of their strengths and weaknesses.

Text Response
Digital Palomar Sky Survey (DPOSS): Good coverage of the sky but scans of photographic plates had undersampled resolution. NVO copies of data from Kitt Peak: Processing can't handle images taken at rates of moving objects and pipeline is too slow for studies of fast moving asteroids.

MAST. Has a very efficient search engine and preview capabilities. CDS. Effective search tools including title and position. Data merging of similar data sets.
No longer in astronomy.
IPAC - NED provides very useful overview and aggregation as front end. Spitzer Space Telescope data straightforward (if time-consuming) to retrieve, but source confusion can be an issue. SDSS - complex sql queries require some thought and training; large-scale ones can be tedious. MAST - HST archives very easy to search by target or position; more challenging to collect all relevant calibration data if not using pipeline reduction. CDS - straightforward for what it is.
I use Vizier from CDS quite a lot. It provides access to published data tables and sever large catalogs (e.g. USNO-B1, SDSS), however it doesn't always link to the underlying data (e.g. you get the published table of radial velocities, but can't get the spectra used to measure these).
--SDSS -- relatively robust for searching, professionally curated/maintained --MAST -- relatively robust for searching, professionally curated/maintained --Various coding repositories (svn, hg, git) through private and public repositories (bitbucket, github) -- easy to use, submit, etc., but can be time-consuming to cleanly maintain for many users
NA
NASA/ADS ArXiv SDSS skyserver

Statistic	Value
Total Responses	8

3. 3. What are some possible use cases for a hub at U of A for previously uncurated astronomy data?

Text Response
Spacewatch Project images of the night sky from 1984-present: Safe harbor for data collected with soft money.
Ground based optical/NIR spectroscopy from small facilities that lack an archive. Reduced spectra are small and with good meta-data in their FITS files can effectively be searched.

o Preserving the "long tail" of research data. o Supporting integrity and reproducibility of research. o Supporting re-use of research data for other purposes, thereby enhancing its value.
Support AAS publications by curating the data underlying the figures for the long term. Interesting question about supporting calibrated images from which measurements were made. Support AAS publications with static copies of modeling code and/or simulations. Accept medium-size, uniformly calibrated data sets at owners' request. (Catalina Sky Survey? AZTEC image data? AGES spectra? WOCS spectra?)
I don't know enough about the possible scope of this "hub" to comment specifically.
--Source code/binaries for simulation codes and analysis routines --Specific versions of binaries/parameter files that were used in generating datasets associated with specific publications --Observational datasets used in specific publications
NA
1. Variability studies 2. Proper motions 3. photometry in non-standard bands 4. spectroscopic information ... but depends a *lot* on the ability to ingest well calibrated data sets. Utility of these will become increasing limited with time, given the rise of wide-field survey programs

Statistic	Value
Total Responses	8

4. 4. Do you envision opportunities for education and outreach associated with curation of “dark data” in astronomy? If so, please describe.

Text Response
Only as much as other archives do education and outreach which to me seems to be "not much".
If the data are easy to find and download, then certainly they could be used in education and outreach. To make the data easy to find and download, we need ways to extract critical metadata and make them searchable.
Any systematic time-domain imaging could be suitable for E&O or citizen science - discovery of novae in local galaxies, comets, supernovae, etc.

Any number of educational lessons could be derived, e.g. using a pre-selected dataset show the student how to compute the age of an open cluster, then have them find cluster data and apply the technique, explain differences with accepted values, demonstrate measurement error or data quality principles etc. Similarly, citizen science projects could be developed to mine previously unpublished data, however not all data would be suitable for such an effort.
I suppose there could be a variant on the Zooniverse that could tie into this data hub, but it seems like that would be a lot of additional work. Perhaps a collaboration with the creators of Zooniverse is warranted.
NA
Astronomy "zoo" projects, but hard to do with non-uniform data sets. One example may be proper motions. See Harvard plate stack project - http://tdc-www.harvard.edu/plates/

Statistic	Value
Total Responses	**7**

5. 5. Please list any other thoughts or suggestions for the workshop, including additional topics for discussion, as well as opportunities and challenges associated with this proposed project.

Text Response
A big question is that of sustainability. Who will pay for long-term storage and associated infrastructure for discovery and access?
Given the open-ended problem, development of clear criteria for what datasets have the most value for additional use would be an important outcome. (This does not mean guessing about science topic - it relates to quality of calibration, sufficiency of information about external conditions, well described uncertainties, etc.) How much searchability and linkage is desired for data supporting publications outside of the context of the article? It is important to define a clear case for why this activity should not just be incremental at an existing major astronomy repository.
Given fundamental differences in radio, optical, X-Ray, Solar, Planetary, etc data, if everything being considered? What about images versus spectra versus catalogs versus event lists versus ..... ? The breadth of instrument outputs, data file formats, use cases associated with a particular sub-field make a general "data

hub" a very ambitious idea beyond being a simple file store. Similarly, is software included in this scope? Data versioning is another issue that needs to be considered.
There are other efforts to create a data hub on a national scale. I am aware of a group at the National Center for Supercomputing Applications and the National Data Service who are working to build a massive data hub for storage of datasets pan-science. To date, they have a working prototype involving manipulation of data in situ on their servers to avoid too many large data transfers of massive datasets. It may be worthwhile to communicate with them to see if a shared infrastructure is warranted. Two autonomous efforts by the both of us may not be as strong as one where we work together.

Statistic	Value
Total Responses	**4**