Project PHaEDRA: Preserving Harvard's Early Data and Research in Astronomy

Daina Bouquin^{1,*}, Katie Frey¹, Maria McEachern¹, James Damon¹, Daniel Guarracino¹, Alex McGrath¹, Edwin Henneken², and Lindsay Smith Zrull³

¹Wolbach Library, Harvard-Smithsonian Center for Astrophysics, USA.

²SAO/NASA Astrophysics Data System, USA.

³Harvard College Observatory, Astronomical Photographic Plate Collection, USA.

Abstract. The staff of Wolbach Library, in collaboration with partners at both the Smithsonian Institution and Harvard University, has begun a complex digitization and transcription effort aimed at making a large collection of historical astronomy research more findable, accessible, interoperable, and reusable (FAIR). This collection of material was originally produced from the mid-18th century through the early 20th century by researchers at the Harvard College Observatory and was recently re-discovered in the HCO Plate Stacks holdings. The team of professionals supporting the effort to make this century and a half old science FAIR have developed a novel, distributed workflow to ensure that people can engage critically with this material to the fullest extent possible. The project's workflow is guided by the collections as data imperative conceptual frameworks and is now being referred to as Project PHaEDRA, or Preserving Harvard's Early Data and Research in Astronomy.

1 Introduction

A prodigious collection of historical astronomy research produced from the mid-18th century through the early 20th century by researchers at the Harvard College Observatory (HCO) was recently re-discovered after decades of being overlooked in an off-site storage facility owned by Harvard University, the Harvard Depository. This unique collection helps document the history of women's contributions to astronomy, the history of HCO, and also comprises remarkable examples of primary source material showing the evolution of observation methods and early astronomy as a whole. The staff of Wolbach Library at the Harvard-Smithsonian Center for Astrophysics (CfA), recognizing the need to share this collection with the broadest possible community, built a multiskilled team and comprehensive workflow to ensure that the necessary infrastructure was in place to digitize the collection and enrich its metadata to the fullest extent possible using transcription. Throughout the process of defining the workflow, Wolbach was guided by the collections as data imperative conceptual frameworks and worked to meet goals originally developed to support modern open data practices: The FAIR Data Principles (Table 1).

*e-mail: daina.bouquin@cfa.harvard.edu ORCID: 0000-0003-2626-3688

1.1 Project PHaEDRA Goals

The Project PHaEDRA team is working to achieve the following:

1. Digitization of the entire PHaEDRA collection
2. Indexing of the collection in HOLLIS and OASIS (Harvard's library catalog systems) and the NASA/SAO Astrophysics Data System (ADS) [1]
3. Transcription of as much of the collection as possible using the Smithsonian Transcription Center [2] citizen science platform
 - Define a markup procedure to tag instances of sketches, calculations, and data tables
 - Subset the sketch collection – Astronomers and scholars in the History of Science can then help contextualize sketches and assign alt text to the images for increased accessibility
 - Subset calculations and data tables for further metadata enrichment and potential future projects
4. Incorporation of transcriptions into machine-readable metadata to enable full text search capability on ADS

1.2 The PHaEDRA Collection

The materials that make up the PHaEDRA collection include 2,518 journals and logbooks, as well as numerous tip-ins, and loose papers. These items are spread across 118 page boxes. The earliest material the Project PHaEDRA team has encountered in the collection thus far is from 1750. Due to size of the collection and initially sparse metadata, the team has not yet fully determined the more recent end of the date range covered by the collection. It is important to note, however, that a sizable portion of the collection is hand written work of the "Harvard Computers".

The Harvard Computers were women who were predominantly hired by Edward C. Pickering (director of HCO, 1877–1919) to process astronomical data. During their time at HCO, the Computers studied over 130 years of the night sky preserved on glass plate photographs. These women cataloged stars, identified variables, interpreted stellar spectra, counted galaxies, and measured distances in space. Much of the Computers' work was done using methods they developed themselves and that are still used in modern astronomy and astrophysics. Several of the women even made their own discoveries [3].

In addition to the work of the Harvard Computers, the PHaEDRA collection also contains the early work of William Cranch Bond and his son George Phillips Bond, the first two directors of HCO as well as many other notable astronomers who visited Cambridge in the mid 1800s. Moreover, the collection's volumes are significant in that they contain many examples of hand drawn observations that predate photography of any sort. Observations done by hand are both exceedingly rare and have the potential to inform modern time domain astronomical studies, thus they are of interest to historians of science and contemporary scientists as well. Artists may also be interested in examining the collection for its aesthetic qualities and ties to the art community. For example, the collection contains previously unknown early sketches by the artist and astronomer E.L. Truvelot [4].

This diverse range of potential applications of knowledge gleaned from the PHaEDRA collection emphasizes that the collection is no less valuable than it was when its items were cutting edge science; rather the possibilities associated with the collection's use have grown more complex and varied.

Above: Pickering and Harvard Computers. 13 May 1913. Back row (L to R): Margaret Harwood (far left), Mollie O'Reilly, Edward C. Pickering, Edith Gill, Annie Jump Cannon, Evelyn Leland (behind Cannon), Florence Cushman, Marion Whyte (behind Cushman), Grace Brooks. Front row: Arville Walker, unknown (possibly Johanna Mackie), Alta Carpenter, Mabel Gill, Ida Woods. Image courtesy of the Astronomical Photographic Plate Collection, Harvard College Observatory.

1.3 FAIR Data Principles and the Collections as Data Imperative

The Project PHaEDRA team aims to achieve its goals of ensuring that the PHaEDRA collection is as broadly useful as possible by working toward meeting the standards outlined in the FAIR Data Principles. These principles came about in 2016, when representatives from academia, industry, funding agencies, and scholarly publishers shared them as general guidelines for enhancing people's ability to engage with research data by making them more findable, accessible, interoperable, and reusable by people, as well as by machines [5]. Although the principles were developed in the context of modern scientific research data needs, they are of vital importance to ongoing conversations within the broader library community concerning library collections in regards to what Thomas Padilla calls, "a data imperative" [6]. According to Padilla, collections as data imperative "entails developing the means to help all members of society, across all classes and backgrounds, working within the academy and outside of it to engage critically with the traces of human activity we collect in the fullest manner possible" [6]. The Library of Congress' National Digital Initiatives team, in response to the collections as data conversation, facilitated the development of three "conceptual frames" [6] to help guide thinking through how libraries can participate in a collections as data imperative:

- Generativity – to increase meaning making capacity
- Legibility – to document and convey provenance and possibility

Above: PHaEDRA collection material – previously unknown sketch by E.L. Truvelot, 7 March, 1873

• Creativity – to empower experimentation

The conceptual frames provided by the Library of Congress supplement the FAIR Data Principles in such a way as to allow the Project PHaEDRA digitization and transcription effort the opportunity to more fully foster the conditions for "meaning making" [4] with this historic collection.

2 Project PHaEDRA Workflow

The Project PHaEDRA workflow is distributed across multiple institutions and systems and requires collection materials to be tracked at both the item and box level. To achieve this the PHaEDRA team is developing a database and custom front end to allow all members of the team to interface with both the collection metadata and perform item tracking. The database incorporates all steps of the process outlined in Figure 1. These steps ensure that the FAIR Principles are enacted and are guided by the collections as data imperative conceptual frameworks of "Generativity" and "Legibility". This is to say that the motivation behind the defined steps is to create the cyberinfrastructure necessary to help diverse communities navigate the possibilities that arise during the collection’s use, and that as much information as possible about the collection and the project is accessible to them [6].

2.1 Project Partners

The partnerships engendered by Project PHaEDRA collaboration are proving to be an exceptional example of how the collections as data imperative conceptual framework of "Creativity" is also foundational to the successful implementation of the FAIR Principles. "Pursuing a collections as data

Table 1. Table 1: FAIR Guiding Principles

To be Findable	F1. (meta)data are assigned a globally unique and persistent identifier F2. data are described with rich metadata (defined by R1 below) F3. metadata clearly and explicitly include the identifier of the data it describes F4. (meta)data are registered or indexed in a searchable resource
To be Accessible	A1. (meta)data are retrievable by their identifier using a standardized communications protocol A1.1 the protocol is open, free, and universally implementable A1.2 the protocol allows for an authentication and authorization procedure, where necessary A2. metadata are accessible, even when the data are no longer available
To be Interoperable	I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation. I2. (meta)data use vocabularies that follow FAIR principles I3. (meta)data include qualified references to other (meta)data
To be Reusable	R1. meta(data) are richly described with a plurality of accurate and relevant attributes R1.1. (meta)data are released with a clear and accessible data usage license R1.2. (meta)data are associated with detailed provenance R1.3. (meta)data meet domain-relevant community standards

imperative requires creative thinking" and for creative thinking to occur, it is necessary to create a team that "administratively and programmatically... encourages a wide range of experimentation" [6]. The team brought together in PHaEDRA’s workflow have been empowered to create a protocol that builds on the benefits afforded to them through the collaboration between the Smithsonian Institution and Harvard University that makes up the Harvard-Smithsonian Center for Astrophysics (CfA). The CfA collaboration was formalized in 1973 to coordinate the related research activities of the two observatories (the Harvard College Observatory and the Smithsonian Astrophysical Observatory) and to share resources to the two institutions’ mutual benefit. Therefore, the partners on the project already have access to a set of administrative and programmatic infrastructures to support their collaboration and have been able to develop completely new ways to take advantage of it.

The partners and their roles:

1. Wolbach Library, Harvard-Smithsonian Center for Astrophysics

- Coordinating all project partners and efforts
- Defining procedures for tracking both physical and digital objects
- Determining metadata procedures and standards
- Prioritizing batches of material for digitization and transcription
- Engaging with the public in transcription efforts
- Planning for long term curation

2. Harvard College Observatory Plate Stacks

- Providing historical metadata and references
- Supporting outreach and transcription workflows

3. NASA/SAO Astrophysics Data System

- Hosting of digitized images of collection
- Helping define metadata and file structure needs
- Creating indexed records containing all available metadata (e.g. text resulting from transcription enabling full-text search)
- Minting DOIs and bibcodes for collection objects

4. Harvard Library Preservation and Imaging Services

- Providing collections care and preparation for scanning
- Managing all elements of imaging workflow
- Creating files meeting needed metadata standards and file structures for ADS image hosting ingest procedures

5. Smithsonian Transcription Center

- Providing a citizen science platform and interface for the public to engage with the digitized material
- Promoting the PHaEDRA project
- Providing structured metadata with transcription files
- Allowing Wolbach to define markup procedure for transcription volunteers ("volunpeers")

6. Harvard University Archives

- Providing access to and support for ArchivesSpace – the collection's finding aid platform
- Offering guidance on archival metadata practices (EAD)

2.2 Workflow Narrative

The general procedure for PHaEDRA workflow is as follows:

Wolbach Library recalls a small batch of material (e.g. three boxes) from off-site storage. Wolbach examines the material and adds as much metadata as possible from the items using the Library's tracking database. The metadata will subsequently be added to the item records in the collection finding aid on OASIS and in ADS records. Next, the boxes are sent to Preservation and Imaging Services where they are prepped and safely scanned. Imaging Services sends the boxes back to Wolbach with an accompanying physical hard drive containing the image files in a predefined structure. Wolbach receives the boxes and hard drive and returns the boxes to off-site storage. Wolbach then gives the hard drive to the NASA/SAO Astrophysics Data System along with access to a spreadsheet containing the collection's item-level metadata. ADS creates records for each item in the batch and hosts the collection image files on their own servers. ADS mints appropriate unique identifiers and creates persistent links. ADS shares the resulting records with Wolbach, which are added to the OASIS record and finding aid. Wolbach then generates a list of URIs for the images that ADS is now hosting. Wolbach provides that list of URIs to the Smithsonian Transcription Center, which uses them

to populate "projects" (one for each collection item) on its platform. The public can then transcribe and review transcriptions of the images on the Transcription Center site even though the image files are being hosted by ADS. Once a "project" is transcribed completely (e.g. one notebook) Wolbach can download the resulting transcription file to be ingested by ADS. ADS can then use their existing procedures to enable full-text search of the item via their public facing search interface.

3 Metadata

The PHaEDRA workflow relies on multiple metadata structures to ensure that the collection's digital objects meet the standards outlined in the FAIR Principles. No proprietary file formats are used and the pipelines rely on open source languages and protocols. More specifically, the workflow employs Python, SQL, CSV, TIF, and EAD at various points in the process. The PHaEDRA finding aid is managed using ArchivesSpace.

3.1 Initial Catalog and Finding Aid

Initially, the Project PHaEDRA team prepared to start from scratch in documenting the collection contents as exceptionally little metadata actually existed in a machine-readable format. The information that did exist was generally at the box level and was abruptly truncated, likely in the process of adding records to the earliest library computer systems at Harvard. The lack of metadata was also likely a product of the collection being relocated on a number of occasions and eventually being left to sit in the Harvard Depository. Luckily, a hand-written item-level catalog documenting the material was serendipitously discovered at the CfA. The hand-written catalog was produced in 1973 by a person named Joseph Timko and was later typed up in 1975. The type-written catalog was discovered soon after the initial catalog was found. Much to the surprise and relief of the PHaEDRA team, a little more searching unearthed a spreadsheet that was created in 1999 based on the Timko catalog. The discovery of the spreadsheet sped up the project exponentially and became the basis on Wolbach Library's work to create a collection finding aid [7].

Wolbach librarians, in consultation with the Harvard Archives, cleaned up the spreadsheet and normalized all dates, names, and existing metadata to the greatest extent possible so that the spreadsheet could be converted into EAD. Python scripts were used to create the EAD and the resulting file was ingested into ArchivesSpace to create the collection finding aid. This data was also shared with the ADS team who used this same information to create item records and mint DOIs and bibcodes for the digitized materials.

Wolbach assigned the following metadata elements to each digital object whenever that information was available:

- id – unique ID defined by Wolbach Library (e.g. phaedra0001)
- kg call number – call number that was likely assigned to items when they were housed in the New England Deposit Library [8] (possibly in the 1950s)
- secondID – a sequence identifier that was applied by the researchers themselves
- titleIdentifier – A non-unique title to describe each item
- yearStart – Earliest date available (year)
- monthStart – Earliest date available (month)
- dayStart – Earliest date available (day)
- yearEnd – Last date available (year)

- monthEnd – Last date available (month)
- dayEnd – Last date available (day)
- boxNumber – Number for the page box the item is housed in
- firstAuthor – A primary author is assigned whenever one can be determined
- secondAuthors – Any other contributors identified in the material
- notes – Content notes gathered in initial examination of the material
- date certainty – Dates are sometimes speculative

The PHaEDRA team will be continuing to flesh out the collection's metadata as it is transcribed and explored more fully.

3.2 Transcription Metadata

The metadata provided through the Smithsonian Transcription Center (STC) emphasizes Project PHaEDRA's collections as data imperative particularly well in that the STC "seeks to engage the public in making our collections more accessible" by working "hand-in-hand with digital volunteers to transcribe historic documents and collection records to facilitate research and excite the learning in everyone" [2]. The STC's focus on engaging actively with the public in meaning-making and the creation of digital objects is essential to the Project PHaEDRA mission to make the collection more FAIR. The STC gives the PHaEDRA team the opportunity to enrich the collection metadata to the point where full-text search is possible. It will also be possible to assign additional meaning to the digitized material thanks to the markup being provided by volunteers. Wolbach worked with the STC to augment their standard instructions for transcribers to ensure that sketches, equations, and data tables were transcribed uniformly, which will allow for them to be more flexibly used as the project progresses [9].

4 Obstacles and Future Work

Despite the Project PHaEDRA team's successful implementation of the project's general workflow, a number of obstacles place limitations on the speed of project implementation and the comprehensiveness of the collection's metadata. Wolbach has begun a social media campaign to adjust to these challenges, but barriers still exist.

4.1 Obstacles

1. The transcription component of the current workflow relies on "volunteers", making outreach and engagement paramount. This engagement defines the project timeline.
2. Some logbooks were initially scanned as part of a separate digitization effort [10] at the CfA and the resulting images were hosted and transcribed separately. These files now need to be incorporated into the new comprehensive workflow.
3. Multi-institution and multi-platform workflow. Communication and documentation are essential and changes to the workflow would be exceedingly challenging
4. Wolbach is receiving only a small amount of funding from the Smithsonian Astrophysical Observatory to scan the material. Funding for physical conservation of materials and support from other potential partners will likely need to be secured externally.

5. Workflows still need to be defined for the sub-setting of images for annotation
6. Approaches must be developed to enable the machine-readability of transcribed data tables

4.2 Potential Future work

The Project PHaEDRA team is confident that they can achieve their initial goals and hope to continue increasing the usefulness of the digital objects that they are creating. One potential avenue of support for that objective would be to link digitized journals to their corresponding DASCH plates. The DASCH project, or Digital Access to a Sky Century @ Harvard [10], is an effort to digitize the entire Harvard College Observatory Glass Plate Collection. The plates themselves are numbered and the PHaEDRA transcriptions will contain those same numbers. Additionally, it may be possible to employ multidimensional image processing techniques to identify sketches and tables without the need for volunteers to tag them, which would dramatically speed up the plan to subset those sketches so that alt text can be added to them. To ensure that the alt text is accurate, the PHaEDRA team is also hoping to develop and fund a post-doc fellowship to contribute to the evolving project and contextualize the materials in the history of astronomy as a whole. Another idea that the team has is to tag the PHaEDRA collection using terms from the Unified Astronomy Thesaurus [11] to ensure that the collection is as findable as possible.

5 Thank you!

Wolbach Library extends a grateful thank you to our partners at the SAO/NASA Astrophysics Data System, the Smithsonian Transcription Center, and Harvard University Library's Imaging Services, Archives, and Preservation Services groups. The team is especially grateful to the many members of the public who have contributed to the transcription of the PHaEDRA material and who will continue to support the project moving forward.

References

- [1] *ADS Search*, <https://ui.adsabs.harvard.edu/>
- [2] *Smithsonian Digital Volunteers*, <https://transcription.si.edu/>
- [3] *Women Computers*, <https://platestacks.cfa.harvard.edu/women-computers>
- [4] *The Trouvelot Astronomical Drawings (1882)*, <https://publicdomainreview.org/collections/the-trouvelot-astronomical-drawings-1882/>
- [5] M.D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.W. Boiten, L.B.d.S. Santos, P.E. Bourne et al., *The FAIR Guiding Principles for scientific data management and stewardship* (2016), <https://www.nature.com/articles/sdata201618>
- [6] T. Padilla, *On a collections as data imperative*, http://www.digitalpreservation.gov/meetings/dcs16/tpadilla_OnaCollectionsasDataImperative_final.pdf
- [7] *Harvard College Observatory observations, logs, instrument readings, and calculations : an inventory*, <http://oasis.lib.harvard.edu/oasis/deliver/deepLink?collection=oasis&uniqueId=wol00001>
- [8] K.D. Metcalf, *The New England Deposit Library* (1942), <http://www.jstor.org/stable/4302994>

- [9] *Women Computers at the Harvard College Observatory* | *Smithsonian Digital Volunteers*,
<https://transcription.si.edu/instructions-women-computers-workbooks>
- [10] *DASCH Home Page*, <http://dasch.rc.fas.harvard.edu/project.php>
- [11] *Unified Astronomy Thesaurus*, <http://astrothesaurus.org/>

Figure 1. Full Project PHaEDRA workflow. HD refers to the Harvard Depository (off-site storage). HDD represents the physical hard drive used to transport the digitized image files. Diagram by Daniel Guarracino