DESIGNING A MACHINE LEARNING LOCAL DATA DICTIONARY. M. N. Le[1] and J. M. McAuley[1], [1]Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA 91109, minh.n.le@jpl.nasa.gov, michael.mcauley@jpl.nasa.gov

Introduction: In order to support the description of data across diverse science domains, the Planetary Data System (PDS) maintains a series of discipline and mission data dictionaries, as well as a broader common dictionary. While current dictionaries support robust descriptions for data products from planetary missions, the PDS ecosystem is lacking in metadata focused on products that are related to, but not directly produced by, mission operations. In particular, there lacks focused metadata to capture the contributions of the machine learning community, whose work often makes use of planetary surface and orbital imagery from PDS collections, especially from the Cartography and Imaging Sciences Node (Imaging Node) archive. The metadata that is available and applicable to machine learning products remain broad and general, leaving much of the contextual data story untold and ultimately hindering future usability of datasets. In order to fill this gap, the Imaging Node team has created a dedicated Machine Learning Local Data Dictionary (LDD) to support the long-term preservation of this community's output.

Justification: The PDS Common Dictionary, derived from the PDS Information Model, contains a common set of broad classes and attributes to define the essential, required components of any product label [1] [2]. Therefore, every single PDS4 label makes use of the PDS Common Dictionary, in addition to any other pertinent discipline or mission dictionaries. While it is possible to describe machine learning data products entirely with the Common Dictionary alone, shoehorning this complex domain into general use categories would ultimately not capture enough context, limiting future reproducibility efforts of the archived work. Additionally, the decision to create a standalone discipline dictionary rather than adding these classes and attributes on to another existing discipline dictionary, allows for more broad use in the machine learning community and across all seven of the PDS domains [3].

Dictionary Design Process:

Domain Knowledge Extraction. The Imaging Node team kicked off initial dictionary design by meeting with subject matter experts in the machine learning field at the Jet Propulsion Laboratory on a biweekly cadence to extract domain knowledge to better understand the problem space. Publications in the field were also supplied by the team of domain experts, to provide a full account of the processes, tools, and datasets involved in developing a machine learning model. The Imaging Node team then conducted an audit of example machine learning artifacts to develop initial descriptions for the products in the context of the PDS archive. Products such as predicted class outputs, validation, training, and test datasets, and image labeling guides were all examined for reference.

Knowledge Modeling. Following PDS best practices and precedence, an ontology model was then created using the Protégé modeling tool [4] [5]. This ontology served as a high-level, object-oriented view of the domain and as a visual anchor for group discussions and shared understanding of concepts.

XML Implementation. Again, following current PDS convention, the object-oriented model was then translated into Extensible Markup Language (XML), the archive system's default structured language for all metadata labels. The dictionary file itself was maintained in a Github repository to allow for ease of collaboration and full tracking of changes.

Iteration and Community Review. Machine learning subject matter experts and senior PDS metadata professionals provided consistent feedback throughout the iterative development process of the Machine Learning LDD. The dictionary team continued biweekly tag-ups with machine learning colleagues, workshopping not only definitions for concepts, but also term relationships, rules, and data properties (data type, cardinality, and nullability). Members of the PDS4 Data Design Working Group (DDWG) were also consulted for guidance, and upon official completion, the dictionary will additionally undergo testing and validation before official release to the public.

Use Case: The dictionary's inaugural use will center on preserving machine learning models used in Mars image content classification work, conducted by Kiri Wagstaff et. al [6]. With this Machine Learning LDD, the Imaging Node team will be able to capture the relationships between the many products involved in developing a model for image content classification, providing users with procedural context. The dictionary will be able to capture details such as the types of algorithms used, parameter settings specified, specific pointers to the training, test, and validation sets that informed the model, and predictions generated by the model.

Conclusion: The ability to fully describe machine learning work for long-term archiving is significant. Beyond furthering research within the machine learning domain itself, this community's work also highlights a unique use of PDS archival holdings and inclusion of their products will further diversify the archive's overall content. The Imaging Node team

hopes that the creation of this dictionary will ultimately encourage widespread archival practices in this field for posterity.

Acknowledgments: This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. © 2021. California Institute of Technology. Government sponsorship acknowledged.

The authors wish to acknowledge Kiri Wagstaff, Steven Lu, Jake Lee, and Paul Horton for their subject matter expertise and continued collaboration on this dictionary effort. Additionally, the authors acknowledge the PDS Engineering Node and members of the PDS4 Data Design Working Group for their guidance and support.

References: [1] PDS Common Dictionary. https://pds.nasa.gov/datastandards/dictionaries/index-1.15.0.0.shtml#pds-common.

[2] PDS4 Information Model Specification. https://pds.nasa.gov/datastandards/documents/im/current/index_1F00.html.

[3] About the PDS. https://pds.nasa.gov/home/about.

[4] Hughes, J. S., Crichton, D., Hardman, S., Law, E., Joyner, R., & Ramirez, P. (2014). PDS4: A model-driven planetary science data architecture for long-term preservation. In *2014 IEEE 30th International Conference on Data Engineering Workshops* (p. 134-141).

[5] Protégé. https://protege.stanford.edu.

[6] Wagstaff, K., Lu, S., Dunkel, E., Grimes, K., Zhao, B., Cai, J., ... & Mandrake, L. (2021). Mars Image Content Classification: Three Years of NASA Deployment and Recent Advances. In *Proceedings of the Thirty-Third Annual Conference on Innovative Applications of Artificial Intelligence*.

[7] Guinness, E. (2018) PDS4 Local Data Dictionaries, *Europa Clipper PDS4 Training Session*.