THE DATA CITATION COMMUNITY OF PRACTICE: AN INTRODUCTION AND REPORT. C. M. Coward[1], D. A. Agarwal[2], and K. A. Copas[3] [1]Jet Propulsion Laboratory, California Institute of Technology (4800 Oak Grove Drive, M/S 111-113, Pasadena, CA, 91109 caroline.m.coward@jpl.nasa.gov) [2]Lawrence Berkeley National Laboratory (1 Cyclotron Road, MS 50B-2239, Berkeley, CA 94720, daagarwal@lbl.gov) [3]Global Biodiversity Information Facility (Universitetsparken 15, DK-2100 Copenhagen Ø, Denmark, kcopas@gbif.org).

Introduction: The open data movement has created a need to accurately cite data sets in the scientific literature, due to citation requirements from journal editors and publishers. However, at present there is no one standard citation for these data sets; Individual publishers have developed their own data citation standards, but these vary widely across the publishing industry.

Data set citations may be found in a variety of sections in a paper, including the acknowledgements, bibliography, abstract, or the body of the article. Similarly, how an author cites a data set varies widely as well; For example, an author may spell out the name of the repository where the data resides, they may use only the acronym, they may cite the repository manager's name, or they may simply mention the larger organization that hosts the repository. These variations not only prove frustrating to the author, who must comply with an ever-increasing list of publisher citation requirements, for the reader who must decipher an infinite variety of citation possibilities, but also for the repository manager, who often must spend hours, weeks, and months combing through the relevant literature to tease out citations from their repository, in order to show the level of usage to repository stakeholders and funders.

Currently, there is no reliable way to apply machine learning to find data citations in the literature. Either the search algorithm is tuned broadly enough to return a large number of results, but with a low level of accuracy (too many "false" results), or the algorithm is tuned more precisely to return a smaller number of results, but missing too many "true" results. New structures, such as dynamic digital object identifiers (DOIs), with multiple data sets or even formats add to the complexity of the environment. [1]

The Community: This issue was brought up both formally and informally at a series of data-focused sessions at a number of conferences in late 2020, primarily the American Geophysical Union 2020 Fall Meeting (held virtually). Comments both in the online chat and verbally indicated that the frustration with data citations is global and widespread. Over the course of the meeting, commenters agreed to form a semi-formal community of practice, to uncover "sticky problems" of all sorts around citing data sets in the literature. The comments generally fell into a loose framework around these challenges. The framework hangs on five main areas:

1. Credit
2. Interconnection
3. Stability
4. Common Approach
5. Culture

To date, approximately 90 people worldwide – planetary scientists, data scientists, repository managers, information scientists, librarians, and others – have stepped forward to participate in the conversation, to be involved, and to potentially solve this issue.

The leaders of the COP (the authors of this abstract and others) proposed several ways to move the conversation forward, including:

- Workshops, webinars, and other professional education events where the Community can work through data citation use cases that are still problematic
- Maintaining a Slack, Discord, or other group communication forum, where members can schedule synchronous conversations
- Generating a body of scholarly work by Community members, including journal publications and session proposals at major science conferences

As a first workshop topic, Deb Agarwal of Lawrence Berkeley Lab proposed a "sticky problem" – how to cite "nested" or "dynamic" DOIs, where a single DOI refers to a curated collection of multiple data sets. On April 8, 2021, the COP leaders hosted the first workshop on this topic. The workshop attracted over 100 attendees, who engaged in spirited conversations about and around the topic. [2]

Next Steps: The Data Citation Community Of Practice plans to have several more workshops and guided discussions on other "sticky problems", as well as conference presentations and journal publications on more targeted instances of internal issues, in order to develop a body of professional literature around the topic of data citations.

Acknowledgments: The authors wish to thank Shelley Stall and Christopher Erdmann of AGU for their leadership in the Community and contributing many hours organizing and providing technical expertise to this initiative. The authors would also like to dedicate

this briefing to the late Dr. Peter Fox of Rensselaer Polytechnic Institute, who, as a leader in the field, provided much energy behind this topic, with many thought provoking comments and much-needed alternative viewpoints.

References:

[1] Vannan, S., et al. (2020), Eos 101, https://doi.org/10.1029/2020EO151665. [2] Agarwal, D. et al. (2021, April). Data Citation Community of Practice. http://doi.org/10.5281/zenodo.4673622.

Additional Information: If you would like to join the Data Citation Community of Practice, please contact Caroline Coward at caroline.m.coward@jpl.nasa.gov