Upstream

INFORMATE: When Are the Data?

Ted Habermann, **Jamaica Jones**, **Howard Ratner** and **Tara Packer**

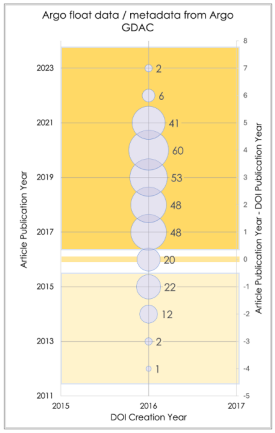
Published February 13, 2024

Citation

Habermann, T., Jones, J., Ratner, H., & Packer, T. (2024). INFORMATE: When Are the Data?. *Upstream*. https://doi.org/10.54900/08pke-hyy45

Keywords

Original Research



Data ***reuse*** happens when papers are published after the data, i.e., publishing a paper and connecting it to the data after the data has a DOI. In this case, 258 papers (82%) that referenced the Argo Data were published after 2016.

Creation of the dataset and article DOIs during the same year is termed ***curation***, metadata creation while the dataset is being created. Twenty papers (6%) that referenced the Argo Data were published during 2016.

Re-curation of the article or dataset metadata, i.e., creating a connection by adding a link to the dataset or article metadata, mostly happens when a dataset is created after an article is published. Thirty-seven papers (12%) that referenced the Argo Data were published before 2016

Acknowledgments

Copyright

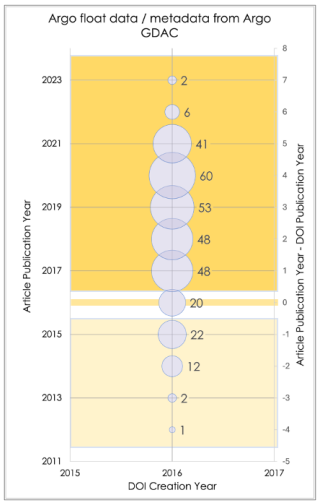
Copyright © Ted Habermann et al. 2024. Distributed under the terms of the Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Upstream

In a recent Upstream blog post we explored where data connected to papers funded by several U.S. Federal Agencies are published. Different data sharing practices across these agencies led to very different distributions of datasets across various repositories. We used CHORUS reports that combine linked article and dataset metadata as input for that work.

The links provided by CHORUS also facilitate comparisons of publication dates for articles and datasets. The Argo Float dataset (10.17882/42182) was cited by over 300 articles funded by NSF between 2012 and 2023. This is a small subset of all articles that use this dataset (Argo Publications), but it provides an example of how comparing publication dates can shed light on temporal relationships of articles and datasets.

Figure 1 compares publication dates for the articles and the creation of the Argo Float dataset DOI. The x-axis shows the creation year of the data DOI (2016). The y-axis show the publication year of the articles (2011 to 2023) on the left and the relative timing of the publications on the right. The number of articles / year is shown as bubble size and as text beside the bubbles.



Data ***reuse*** happens when papers are published after the data, i.e., publishing a paper and connecting it to the data after the data has a DOI. In this case, 258 papers (82%) that referenced the Argo Data were published after 2016.

Creation of the dataset and article DOIs during the same year is termed ***curation***, metadata creation while the dataset is being created. Twenty papers (6%) that referenced the Argo Data were published during 2016.

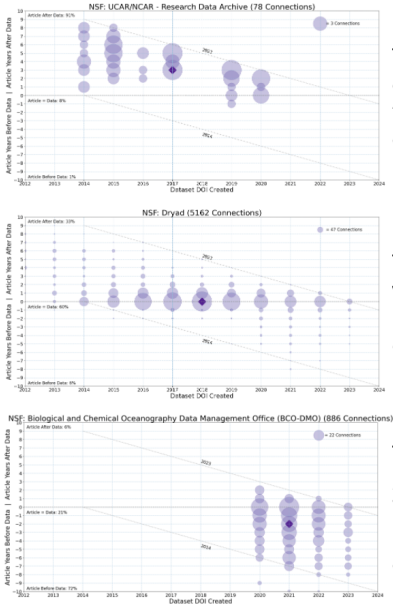
Re-curation of the article or dataset metadata, i.e., creating a connection by adding a link to the dataset or article metadata, mostly happens when a dataset is created after an article is published. Thirty-seven papers (12%) that referenced the Argo Data were published before 2016

Figure 1. Argo Float dataset and article timing. The temporal relationship between published articles and connected datasets defines three types of actions: data reuse, data curation, and data re-curation.

Figure 1 suggests a nomenclature of three terms that may be helpful for describing these temporal relationships and related repository behaviors. First, many repositories are focusing significant efforts on facilitating *reuse* of data they steward. Reuse success appears as bubbles above the centerline in Figure 1, i.e. articles published after datasets. Second, many scientific journals and funders are requiring submission of datasets to repositories as papers are published and structured references to those datasets (Stall et al., 2023). This practice, termed *curation*, appears as bubbles on the centerline of Figure 1, i.e. article publication date = dataset publication date. Finally, many domain repositories are extracting data from past articles and publishing it with identifiers. Connecting these past articles to the newly published data requires adding a connection into the metadata for the paper or into a link repository like ScholeXplorer. This metadata augmentation is a *re-curation* process.

Upstream

This same visualization can be created for multiple datasets in a repository. Examples of these visualizations, termed Connection Timelines, for repositories that focus on re-use, curation, and re-curation are shown in Figure 2.



The Research Data Archive ar UCAR/NCAR is a good example of a repository with data re-use. Ninety-one% of the articles funded by NSF using data from this repository were published after the datasets were published.

The Dryad Data Repository is a good example of a repository with articles and datasets published at the same time (curation). This reflects the focus of Dryad on publishing datasets that are related to published articles.

The Biological and Chemical Oceanography Data Management Office (BCO-DMO) at Woods Hole is a good example of a repository with articles published before datasets. This requires additions to metadata for the articles or the links (re-curation) to make the connections.

Figure 2. Connection Timelines for datasets connected to papers funded by NSF in three repositories. The % of connections in each of the three categories are shown along the left edge of the timelines and the median position of all connections for the repository are shown as diamonds.

Making these connections with links from publicly available CHORUS reports currently limits the data to articles funded by specific funders, NSF in this case, so the timelines are only for subsets of datasets in these repositories. In many cases, these subsets include hundreds or even thousands of connections, so they may still reflect general repository behaviors.

Connection timelines for many more repositories, described on this poster and available here, include many examples of these behaviors and many repositories with mixed behaviors. Future work will explore other approaches to finding connections to broaden the examples and improve understanding of data reuse, curation, and re-curation as well as repository behaviors.

Acknowledgments

This work is part of the INFORMATE Project, a partnership between Metadata Game Changers and CHORUS. This work was funded by the U.S. National Science Foundation, award 2134956.

References

Argo. (2024). *Argo float data and metadata from Global Data Assembly Centre (Argo GDAC)* [dataset]. SEANOE. https://doi.org/10.17882/42182

Argo Publications (2023). https://argo.ucsd.edu/outrech/publications/.

Upstream

Habermann, T., & Robinson, E. (2024). *What Came First: The Paper Or the Data* (p. 6245187 Bytes). ESIP. https://doi.org/10.6084/M9.FIGSHARE.25139354.V1

Habermann, T., Jones, J., Ratner, H., & Packer, T. (2023). *INFORMATE: Where Are the Data?* https://doi.org/10.54900/vnevh-vaw22

Stall, S., Bilder, G., Cannon, M., Chue Hong, N., Edmunds, S., Erdmann, C. C., Evans, M., Farmer, R., Feeney, P., Friedman, M., Giampoala, M., Hanson, R. B., Harrison, M., Karaiskos, D., Katz, D. S., Letizia, V., Lizzi, V., MacCallum, C., Muench, A., … Clark, T. (2023). Journal Production Guidance for Software and Data Citations. *Scientific Data*, *10*(1), 656. https://doi.org/10.1038/s41597-023-02491-7