Canadian Science Publishing

TOPICAL COMMUNICATIONS

# Prototyping a collaborative data curation service for coastal science[1]

Evan B. Goldstein, Anna E. Braswell, and Caitlin M. McShane

**Abstract:** The growing push for open data resulted in an abundance of data for coastal researchers, which can lead to problems for individual researchers related to data discoverability. One solution is to explicitly develop services for coastal researchers to help curate data for discovery, hosting discussions around reuse, community building, and finding collaborators. To develop the idea of a coastal data curation service, we investigate aspects of the UNESCO International Coastal Atlas Network member sites that could be used to build a curation service. We develop a minimal example of a coastal data curation service, deploy this as a website, and describe the next steps to move beyond the prototype phase. We envision a coastal data curation service as a way to cultivate a community focused on coastal data discovery and reuse.

*Key words:* coastal data, curation, coastal researchers, data discoverability.



## Motivation

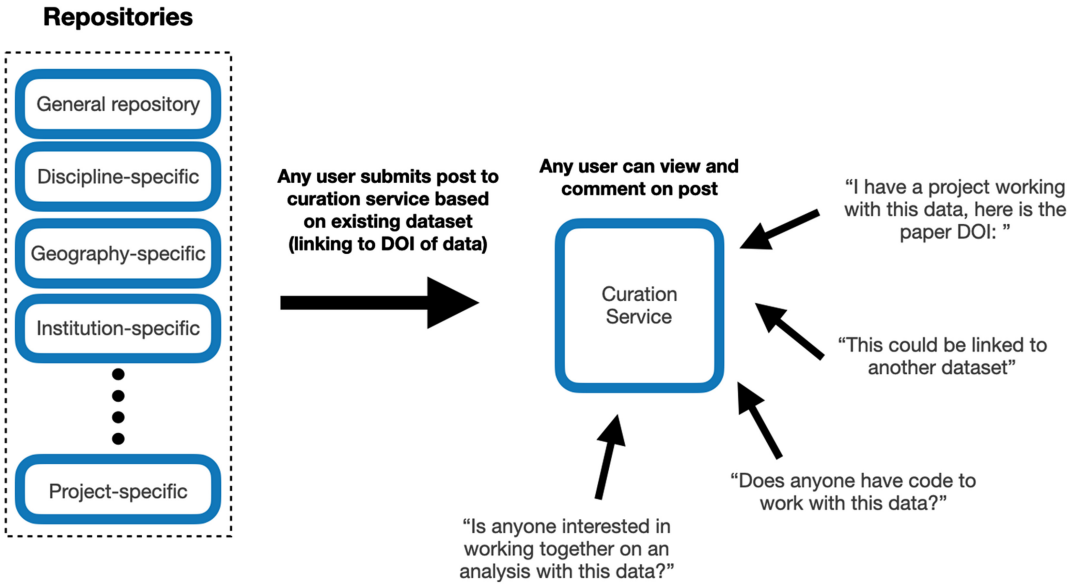
There is a broad push from scientists, scientific societies, publishers, and funders in the Earth and Environmental Sciences for findable, accessible, interoperable, and reusable data (FAIR data; Wilkinson et al. 2016) (e.g., Stall et al. 2018, 2019). This growing volume of open data available for coastal researchers enables data-driven investigations and large-scale data analysis to develop insight into complex coastal issues, which often lie at the intersection of physical, social, and biological processes. A growing compendium of data presents new challenges, one example being discovery — how does a researcher discover if new or useful data already exists? Just as it is challenging to keep up with published literature, it is a challenge to keep up with available and published data.

Relevant data is stored in a variety of places: disciplinary-specific repositories, location-specific repositories, institutional repositories, funder repositories, on personal/project websites, or in non-public facing places (i.e., laboratory/office hard drives as dark data; Heidorn 2008). The existence of a specific data repository may not be known to a researcher who could benefit from data contained within it — i.e., individual researchers must discover a repository to use open data. The large number of repositories listed by

Received 15 March 2021. Accepted 2 June 2021.

**E.B. Goldstein.** Department of Geography, Environment, and Sustainability, University of North Carolina at Greensboro, Greensboro, NC 27412, USA.
**A.E. Braswell.** Fisheries and Aquatic Sciences Program, School of Forest Resources and Conservation, University of Florida/IFAS, Gainesville, FL 32603, USA.
**C.M. McShane.** Department of Geography, University of Colorado Boulder, Boulder, CO 80301, USA.
**Corresponding author:** Evan B. Goldstein (email: ebgoldst@uncg.edu; Twitter: @ebgoldstein).
[1]This paper is part of a Collection entitled: The emerging role of data-driven science at the coast.

**Fig. 1.** Schematic for repository and curation interface.




the re3data.org project (Pampel et al. 2013) is a testament to the accelerated growth in a number of repositories. Stand-alone data descriptor articles aid in dataset discovery, published in venues such as *Earth System Science Data* (Pfeiffenberger and Carlson 2011), *Earth and Space Science* (Hanson 2014), *Geoscience Data* (Allan 2014), and *Scientific Data* (Scientific Data 2014). Frequently, data arrives without its own descriptor article, is discussed and reported in individual manuscripts, and may even be present and found solely in the manuscript or in supplementary material.

The growing library of published data, in a variety of repositories, is analogous to the rapid growth in the number of scholarly manuscripts (e.g., Bornmann and Mutz 2015) in a growing number of journals (e.g., Tenopir and King 2014). This issue is well known in coastal research, where relevant literature spans disciplinary boundaries and is published in an almost dizzying array of venues split by discipline, prestige, geographic location, language, and intended audience. Researchers must keep track of an expanding number of relevant publications per unit time, as well as a growing list of journals, echoing the age-old problem of working to understand the flood of scholarly works (e.g., Blair 2010). Discovery of new research has become a significant time commitment for scholars (e.g., Priem 2013). The challenge of discovery has motivated the development of new services and platforms — new search engines (e.g., Google Scholar), social networks (e.g., ResearchGate), recommendation services (e.g., Shetty et al. 2021), and curation services (e.g., preLights, Peeriodicals, Peer Community In, Papers with Code). These services are geared mostly toward manuscripts, though some analogous services exist for data alone (e.g., DataCite Search, Google Dataset Search, Data is Plural; Singer-Vine n.d.).

We focus here on envisioning a new data curation service to aggregate relevant published data for specific end-users (in this case, coastal researchers) and act as a communal, interdisciplinary space that intentionally generates discussion on various aspects of data products (Fig. 1). This space is different than a data portal, whose role is focused on harmonizing and presenting data relevant for a given geographic region or topic (e.g., Lazarus et al. 2020). Instead, we imagine a curation service would not host or

harmonize data, instead it would allow community members to contribute and highlight coastal datasets that could be of interest to others. A curation service works to point coastal scientists toward relevant datasets archived in various repositories. Data authors can then still deposit data in any repository, which might be subject to funder mandates, institutional mandates, or preferences based on discoverability, submission ease, cost, discipline, or other factors.

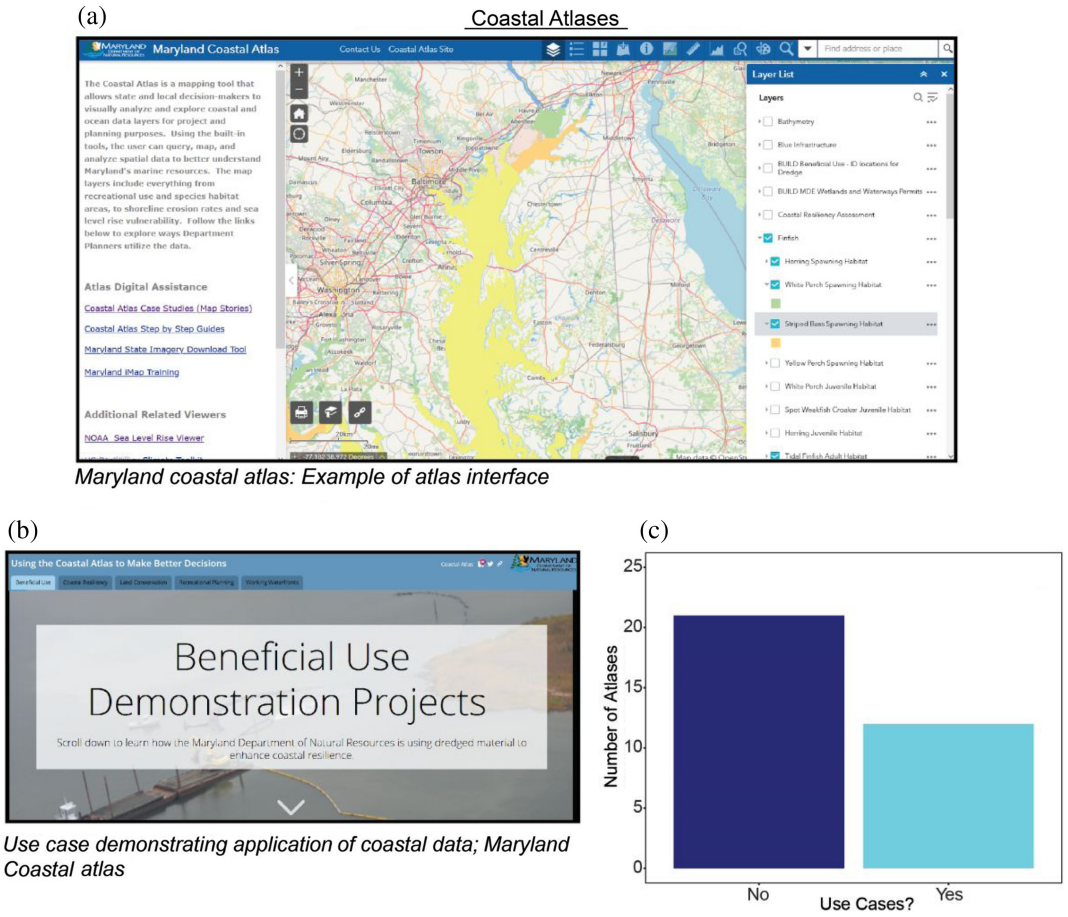
The biggest benefit of a curation service would be to permit interactions around relevant data, such as fostering community discussion around particular datasets and allowing community members to aid in the curation process. We believe hosting discourse around a data set could encourage critical reflection, interaction, and collaboration. Each dataset could also have its own dialogue or forum where people can ask or answer questions pertaining to the data, point out previous uses of the data (e.g., manuscripts, preprints, posters, talks, blogs), link to open-source code for working with specific datasets, and discuss ways that a highlighted dataset can be used in interdisciplinary research. Users could discuss potential uses of data as well as problems encountered while using data. This type of discussion is usually restricted to physical conversations and emails. By developing an open forum for discourse surrounding data usage, specifically tied to the dataset, this approach will further develop a community around coastal data science. An online discussion platform could be used by researchers to actively develop ideas to apply data across multiple subject areas and also promote individual datasets to the coastal community — not just to academic coastal researchers, but to practitioners, managers, consultants, and others working in the coastal realm. In the following sections we investigate online networks that have attempted to coalesce data for the coastal science community. Through this investigation we highlight the pitfalls and successes of prominent data service providers in the coastal domain — Coastal Atlases. We use our findings to propose a flexible lightweight data curation service that addresses these issues and offers a platform to enhance data visibility and foster community discussion.

## UNESCO International Coastal Atlas Network (ICAN)

To understand the features and functions of a new coastal data curation service, we examine an existing network of coastal data providers. Our goal is to understand what data curation services already exist, how they are operationalized, how they differ from a traditional data portal, and any potential problems to be avoided. We examine the UNESCO International Coastal Atlas Network (ICAN), a group of organizations that aims to increase sharing of data relevant to coastal or marine processes and to establish global-level data interoperability (Wright et al. 2011). Member sites serve as both a curation service — presenting a selection of data specific to a location or geography — and a data portal, hosting and potentially harmonizing data. Through their members' locational data curation, ICAN works to expand data interoperability and data sharing by creating a set of best practices for atlas establishment, such as utilizing a single, unified (but proprietary) mapping software. We investigate 31 ICAN member sites and 11 non-member atlas sites to examine features and functions that are relevant for curation services (Fig. 2; Supplementary Table S1[1]). ICAN members seem to individually collect and curate coastal data that are then typically stored in their own large online repositories or as a series of external links designed to take you to the homepage of the institution housing the data product. Many of the individual ICAN members host some of the same datasets, particularly NOAA- or USGS-created datasets.

[1]Supplementary data are available with the article at https://doi.org/10.1139/anc-2021-0002.

**Fig. 2.** Example of a Coastal Atlas. A third of atlases provide example use cases for the data.




We specifically focus on the role of ICAN members as data curators — their function as pointers to datasets that they do not host. With this function come problems that affect the functionality of the atlas. Many datasets on the ICAN member sites do not use persistent identifiers (e.g., Digital Object Identifiers; DOIs) to link to curated, non-hosted data. This problem is not surprising, as many datasets do not have DOIs or any other type of persistent identifier (Goldstein et al. 2017). The lack of any persistent identifier results in datasets across numerous atlases suffering from link rot issues — the web addresses no longer resolve, or no longer link to the original dataset. A key aspect for data curation services, since they are not hosts, is to make sure that the datasets discussed are always available via a persistent identifier. Researchers can then programmatically access data using the DOI or identifier (and appropriate retrieval techniques for individual repositories).

Some coastal atlases do perform a few of the functions of a data curation service. Well-designed atlases (i.e., the Maryland Coastal Atlas: https://dnr.maryland.gov/ccs/coastalatlas/Pages/default.aspx) highlight data through applications such as ESRI ArcGIS Story Maps or blog posts. These added pieces help users to understand the development and possible uses of open coastal data. Although coastal atlases are typically based around geography, they tend to bring data from multiple sources and portals together in one place. This service can help coastal atlas users discover and connect to other coastal data resources. Finally,

coastal atlases often cater to many different user groups, including audiences with varying data literacy. For example, for less experienced users, coastal atlases let users view data of interest on a map through an easy-to-use interface. Often this data is available for download and analysis for the more experienced data user. These three qualities of coastal atlases make data more available and discoverable to broad audiences across the designated geography of the atlas.

## A new prototype curation service for coastal data

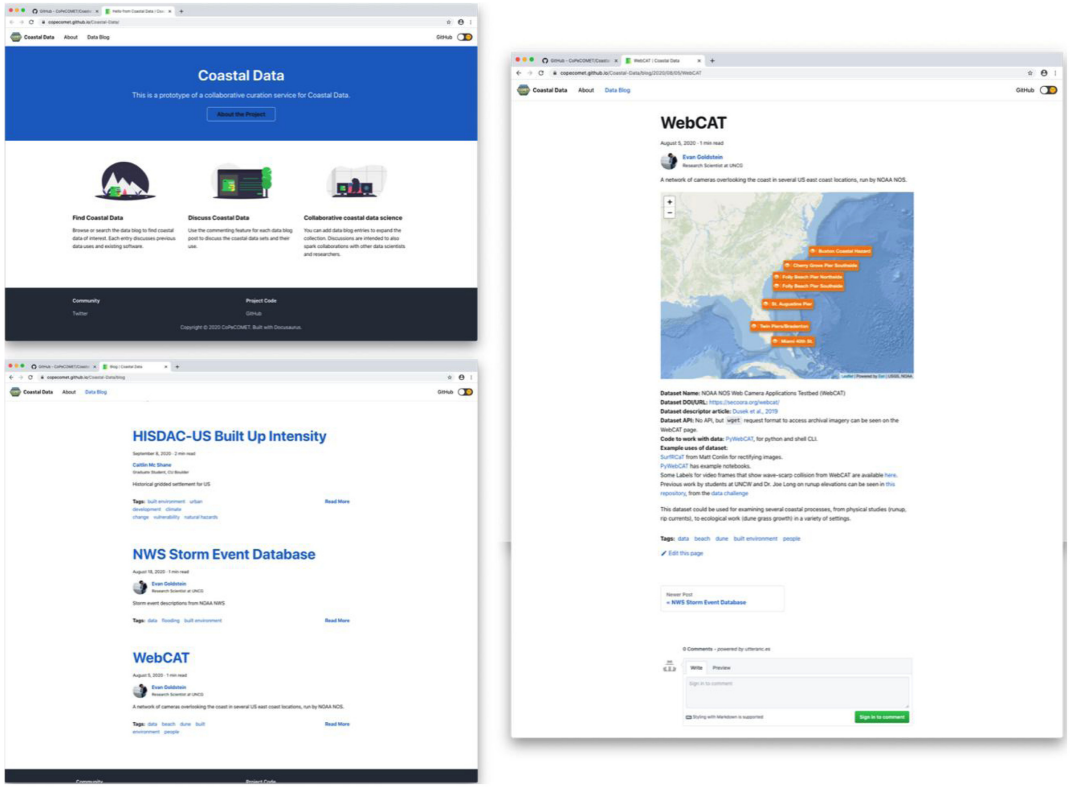
We use the UNESCO coastal atlases as a reference point for developing a service that functions less like a portal and more like a curation service. We envision a new curation service that points researchers toward relevant data but also intentionally fosters community collaboration and data discovery. A service needs to be designed for researchers and end users to easily discover, access, and discuss data products. A minimal example is shown in Fig. 3 and is currently deployed (https://copecomet.github.io/Coastal-Data/) using the static site generator Docusaurus (Facebook 2021). There are three core components. First is an introduction page that describes the goals of the project and how to contribute. Second, individual dataset entries, each with a predefined format. Any user can contribute an entry via adding a markdown document to the public GitHub repository. A template for dataset entries is provided in the repository. Third, a discussions/commenting section that is attached to each post, where anyone can type comments, discussion points, links to code and other relevant info, and flag other users via their GitHub usernames. This system is built on top of GitHub, similar to other new scholarly communication tools (e.g., JOSS; Smith et al. 2018).

Aspects of this curation service model informed by the Coastal atlases, such as introductory pages that describe the scope of the data within, and datasets each with entries based on metadata. Atlases help to potentially combat the overload of datasets being published in different repositories and venues by creating a single source for data information. The difference in our design is that a curation service would explicitly point towards multiple data hosts and is not solely focused on a location or region. As data are not directly stored by the curation service, searching, accessing, and discussing data is a relatively lightweight service that overlays the multitude of repositories currently housing data products. Each entry has a link to the data download page and (or) a DOI that the user can follow if they wish to access data. A link checking algorithm (Lycheeverse 2021), enabled via a GitHub action, verifies each month that links resolve.

This curation service acts as a registry of entries on coastal data products. We have populated our prototype with several examples, but we envision that most entries will be crowd sourced. Each entry contains general information about a dataset and will also include comment sections for caveats, potential data linkages, uncertainty, and help wanted. In addition to general information, data entries include links to publications that have used the data, blog posts, and (or) tutorial pages. To attempt to include researchers with less programmatic ability, links to code snippets that can download, transform, or manipulate data can also be provided on the page. We do not enforce any data format, standards, or interoperability aside from this required metadata. Akin to a preprint server (i.e., EarthArXiv; Narock et al. 2019*a*, 2019*b*) our curation service only has a light moderation policy for entries and data inclusion — that each entry has the requisite metadata and would be of interest to the coastal community.

Discussion threads attached to each data entry can focus on recommendations for other relevant datasets and interdisciplinary discussions about data use. We envision that potential collaborations could establish in these discussion threads to connect scientists to data and other researchers while generating constructive dialogue about data products. As the

**Fig. 3.** Our minimal, prototype example of a deployed curation service. Top left, the splash page. Bottom left, example of dataset entries with tags. Right, example of a dataset entry with links to the original data, examples of code and use of the data.




service builds outward, a service could periodically host guest articles about data trends and novel data products where community members can respond and engage with data scientists from various backgrounds.

Publicizing individual data products will enhance exposure for both the dataset and for potential users and will generate conversation about specific products. Using available data descriptors and published studies, a brief summary of the methodology, validation, uncertainty, and usage will be provided to a discussion forum where community members can ask questions or submit their own anecdotal experience of using that dataset. Community members can rate the data across several categories including ease of use, accuracy, and general applicability.

## Future directions

We have presented and described a minimal example of a coastal data curation service, which is based on our investigation of UNESCO coastal atlases. Our prototype only highlights a few data products currently, but we intend to expand this service by publicizing it and personally contributing more entries. A full-featured coastal data curation service needs to be further developed from this prototype work, but a key to developing a rich experience for participants is actively growing a community of researchers who read and contribute to the effort. We believe that hosting highlighting coastal datasets and hosting discussion surrounding the use of specific datasets could incentivize users to read and then

potentially contribute to the effort — i.e., the service should offer some value to researchers, and hope that those participants then contribute new datasets or pointing new people to the service.

## Code availability

Our minimal example of our curation service is currently deployed at https://copecomet.github.io/Coastal-Data/, and the source code for the site is available via Goldstein et al. (2021) and on GitHub: https://github.com/CoPeCOMET/Coastal-Data.

## Acknowledgements

We thank the editor Ian Townend and three anonymous reviewers for their comments. The authors gratefully acknowledge support from US National Science Foundation (1939954, 1953412 to E.B.G.; 1940006, 1924670 to A.E.B.), a US National Science Foundation Graduate Research Fellowship (Grant No. DGE 1650115 to C.M.M.), and an Early-Career Research Fellowship from the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine (to E.B.G.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the Gulf Research Program of the National Academies of Sciences, Engineering, and Medicine.

## References

Allan, R. 2014. Geoscience data. Geosci. Data J. **1**: 1. doi:10.1002/gdj3.3.
Blair, A.M. 2010. Too much to know: Managing scholarly information before the modern age. Yale University Press.
Bornmann, L., and Mutz, R. 2015. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. J. Assoc. Inf. Sci. Technol. **66**(11): 2215–2222. doi:10.1002/asi.23329.
Facebook. 2021. Docusaurus. GitHub repository. Available from https://github.com/facebook/docusaurus.
Goldstein, E.B., Braswell, A.E., and Mc Shane, C.M. 2021. A prototype coastal data curation service. GitHub repository. Available from https://github.com/CoPeCOMET/Coastal-Data.
Goldstein, J.C., Mayernik, M.S., and Ramapriyan, H.K. 2017. Identifiers for earth science data sets: Where we have been and where we need to go. Data Sci. J. **16**: 23. doi:10.5334/dsj-2017-023.
Hanson, B. 2014. AGU to launch a new open-access journal spanning the earth and space sciences. Eos, Trans. AGU, **95**(6): 56–56. doi:10.1002/2014EO060004.
Heidorn, P.B. 2008. Shedding light on the dark data in the long tail of science. Libr. Trends, **57**(2): 280–299. doi:10.1353/lib.0.0036.
Lazarus, E.D., Aldabet, S., Thompson, C.E.L., Hill, C.T., Nicholls, R.J., French, J.R., et al. 2020. The UK needs an open data portal dedicated to coastal flood and erosion hazard risk and resilience. EarthArXiv. doi:10.31223/X5989C.
Lycheeverse. 2021. lychee-action. GitHub repository. Available from https://github.com/lycheeverse/lychee-action.
Narock, T., Steeve, V., Nüst, D., and Whitehead, B. 2019*a*. eartharxiv/eartharxiv.github.io: EarthArXiv Resources Website (Version v1.1). Zenodo. doi:10.5281/zenodo.2554512.
Narock, T., Goldstein, E.B., Jackson, C., Bubeck, A., Enright, A., Farquharson, J., et al. 2019*b*. Earth science is ready for preprints: The first year of EarthArXiv. Eos, **100**. doi:10.1029/2019EO121347.
Pampel, H., Vierkant, P., Scholze, F., Bertelmann, R., Kindling, M., Klump, J., et al. 2013. Making research data repositories visible: The re3data.org registry. PLoS ONE, **8**(11): e78080. doi:10.1371/journal.pone.0078080. PMID:24223762.
Pfeiffenberger, H., and Carlson, D. 2011. "Earth system science data" (ESSD) — A peer reviewed journal for publication of data. D-Lib Mag. **17**(1/2). doi:10.1045/january2011-pfeiffenberger.
Priem, J. 2013. Beyond the paper. Nature, **495**: 437–440. doi:10.1038/495437a. PMID:23538811.
Scientific Data. 2014. More bang for your byte. Sci. Data, **1**: 140010. doi:10.1038/sdata.2014.10.
Shetty, N., Verstak, A., Hwang, K.J., Jin, L., David, P., and Acharya, A. 2021. Scholar recommendations reloaded! Fresher, more relevant, easier. Available from https://scholar.googleblog.com/2021/02/scholar-recommendations-reloaded.html.
Singer-Vine, J. n.d. Data is plural. Available from https://tinyletter.com/data-is-plural [accessed 9 March 2021].
Smith, A.M., Niemeyer, K.E., Katz, D.S., Barba, L.A., Githinji, G., Gymrek, M., et al. 2018. Journal of Open Source Software (JOSS): Design and first-year review. PeerJ Comput. Sci. **4**: e147. doi:10.7717/peerj-cs.147.
Stall, S., Yarmey, L.R., Boehm, R., Cousijn, H., Cruse, P., Cutcher-Gershenfeld, J., et al. 2018. Advancing FAIR data in earth, space, and environmental science. Eos, **99**. doi:10.1029/2018EO109301.
Stall, S., Yarmey, L., Cutcher-Gershenfeld, J., Hanson, B., Lehnert, K., Nosek, B., et al. 2019. Make scientific data FAIR. Nature, **570**: 27–29. doi:10.1038/d41586-019-01720-7. PMID:31164768.
Tenopir, C., and King, D.W. 2014. The growth of journals publishing. *In* The future of the academic journal. Chandos Publishing. pp. 159–178.

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, **3**: 160018. doi:10.1038/sdata.2016.18. PMID:26978244.

Wright, D.J., Cummins, V., and Dwyer, E. 2011. The international coastal atlas network. *In* Coastal informatics: Web atlas design and implementation. IGI Global. pp. 229–238.