ICES Journal of Marine Science

ICES Journal of Marine Science (2019), 76(6), 1415–1423. doi:10.1093/icesjms/fsz068

Quo Vadimus

Disrupting data sharing for a healthier ocean

Linwood H. Pendleton[1,2,3,4]*, Hawthorne Beyer[3], Estradivari[5], Susan O. Grose[4], Ove Hoegh-Guldberg[3], Denis B. Karcher[4], Emma Kennedy[3], Lyndon Llewellyn[6], Cecile Nys[4], Aurélie Shapiro[1], Rahul Jain[7], Katarzyna Kuc[7], Terry Leatherland[7], Kira O'Hainnin[7], Guillermo Olmedo[7], Lynette Seow[7], and Mick Tarsel[7]

[1]*World Wildlife Fund, 1250 24th Street NW, Washington, DC 20037, USA*
[2]*The Nicholas Institute for Environmental Policy, Duke University, 2117 Campus Drive, P.O. Box 90335, Durham, NC 27708, USA*
[3]*Global Change Institute, Research Road, The University of Queensland, St. Lucia QLD 4072, Australia*
[4]*Ifremer, CNRS, UMR 6308, AMURE, IUEM, University of Western Brittany, Technopôle Brest-Iroise, Rue Dumont D' Urville, Plouzané 29280, France*
[5]*Conservation Science Unit, WWF - Indonesia, Graha Simatupang Tower 2 Unit C, 7th - 11th Floor, Jalan Letjen TB Simatupang Jakarta 12540, Indonesia*
[6]*Australian Institute of Marine Science, Townsville, 1526 Cape Cleveland Road, Cape Cleveland QLD 4810, Australia*
[7]*IBM Corporation, Corporate Citizenship & Corporate Affairs, 2455 South Road, Poughkeepsie, New York 12601, USA*

**Corresponding author: tel: +33 7 82 83 18 23; e-mail: linwood.pendleton@wwf.org.*

Pendleton, L. H., Beyer, H., Estradivari, Grose, S. O., Hoegh-Guldberg, O., Karcher, D. B., Kennedy, E., Llewellyn, L., Nys, C., Shapiro, A., Jain, R., Kuc, K., Leatherland, T., O Hainnin, K., Olmedo, G., Seow, L., and Tarsel, M. Disrupting data sharing for a healthier ocean. – ICES Journal of Marine Science, 76: 1415–1423.

Received 10 January 2019; revised 22 March 2019; accepted 22 March 2019; advance access publication 29 April 2019.

Ocean ecosystems are in decline, yet we also have more ocean data, and more data portals, than ever before. To make effective decisions regarding ocean management, especially in the face of global environmental change, we need to make the best use possible of these data. Yet many data are not shared, are hard to find, and cannot be effectively accessed. We identify three classes of challenges to data sharing and use: uploading, aggregating, and navigating. While tremendous advances have occurred to improve ocean data operability and transparency, the effect has been largely incremental. We propose a suite of both technical and cultural solutions to overcome these challenges including the use of natural language processing, automatic data translation, ledger-based data identifiers, digital community currencies, data impact factors, and social networks as ways of breaking through these barriers. One way to harness these solutions could be a combinatorial machine that embodies both technological and social networking solutions to aggregate ocean data and to allow researchers to discover, navigate, and download data as well as to connect researchers and data users while providing an open-sourced backend for new data tools.

Keywords: combinatorial machine, collaboration, data aggregation, data sharing, data uploading, ocean conservation.

Introduction

Ocean conservation and management are failing to keep pace with our rapidly changing planet (WWF, 2018). Many marine mammals struggle to survive (IUCN, 2018), fish populations continue to decline, and coral reef ecosystems are dying (Hoegh-Guldberg *et al.*, 2018; FAO, 2018; WWF, 2018). Hypoxic zones are expanding (WWF, 2018), and plastics have infiltrated nearly every part of the marine environment (Jamieson *et al.*, 2017; Haward, 2018; Munthe and Jensen, 2018; WWF, 2018). Climate change and local stressors are having profound impacts on ocean social-ecological systems (Poloczanska *et al.*, 2016; Hoegh-Guldberg *et al.*, 2018).

Evidence-based solutions are needed for these ocean-based problems in order to better manage a rapidly changing ocean (Sutherland *et al.*, 2004; Fisher *et al.*, 2014; Science 20, 2019). To develop such solutions, we need to make best use of available

© International Council for the Exploration of the Sea 2019. All rights reserved.
For permissions, please email: journals.permissions@oup.com

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

data to understand the causes and patterns of change to the physical, ecological, and social components of ocean systems. We need data to inform models of ocean change and its effects on ecosystems and people, to evaluate scenarios associated with proposed actions, and to know whether our policies are working (Sutherland *et al.*, 2004; WWF, 2018). The global and dynamic nature of the ocean and its ecosystems means that data acquisition and sharing must happen on an unprecedented scale and at faster rates. Rapid global environmental change means we must constantly reassess and update what we know. Time is of the essence.

Fortunately, we now have more data about more of the ocean than ever before (Visbeck, 2018; WWF, 2018). Remote sensing platforms continuously collect petabytes of earth observation data (e.g. the Landsat and Sentinel programs). Thousands of scientists are hard at work collecting data in the field (IOC-UNESCO, 2017). Scores of online platforms are emerging where scientists can share and access data (see Supplementary Materials for a sample of ocean data portals).

Yet, we still need more data to effectively manage oceans. While satellites, buoys, and other technical approaches have helped chart and monitor the physical and chemical properties of much of the ocean, as much as 90% of the seafloor remains unmapped and unmonitored. Fifteen percent or less of the ocean is as well mapped as the terrestrial surface of the planet (Sandwell *et al.*, 2003). Our lack of understanding and regular monitoring of the biological and human dimensions of the ocean are likely even more data poor. Many habitats, including the deep sea, ocean trenches, ice-bound waters, methane seeps, and even coral reefs remain poorly studied at the global scale. Geographic gaps in biodiversity data are particularly acute for many parts of the global ocean including coastal areas of the Indian Ocean, the southern and eastern Mediterranean Sea, polar seas, and much of the South American coastal ocean (Costello *et al.*, 2010). The proportion of undiscovered marine species is estimated to be as high as 80% (Costello *et al.*, 2010) with many invertebrate taxa being particularly poorly documented and monitored (Costello *et al.*, 2010). Even organisms as large as whales and dolphins remain consistently under-evaluated and monitored; 52% of all IUCN-listed cetaceans are considered as data deficient (Parsons, 2016). Data about many of these places and organisms exist, but hidden in notebooks and laptops, and not available for the new analyses that are needed for ocean management.

One factor limiting our evidence-base for ocean management is that we use just a fraction of the potentially available data (Figure 1). Many ocean data are never shared publicly (Costello, 2009; Kim and Stanton, 2016). Those data that are shared may be difficult to find and integrate with other datasets. Online platforms are often discipline-specific or application specific, creating barriers to discovery and integration (Arzberger *et al.*, 2004; Chavan and Ingwersen, 2009; Costello, 2009; Kim and Zhang, 2015). Perhaps most challenging, data are easily dissociated from the people who helped create and curate them, rendering communication between users and producers difficult (Ferguson *et al.*, 2014).

Another factor limiting the evidence base that underpins ocean management is that our science and analysis is often limited to "good data," i.e. high quality, in a format that meets some agreed standard. But much information exists in the messy data that may be of varying formats and quality, infrequently, if ever, updated and hence easily corrupted, changed, or simply lost (Costello, 2009). Some of these messy data are rigorously collected by students, non-academic researchers, academic researchers that do not publish their data regularly, government scientists, and others but may not be perceived to be "good data" owing to a lack of standardization, a reliance on technology of unknown accuracy or other factors that may lead one to question the rigour of the data. Other data may fall more in the realm of citizen science or even big data that are collected passively from non-scientific sources (e.g. Instagram™, Flicker™, and other types of social media). While messy, these data may still contain important information, especially where large gaps in "good data" exist. We need to find new ways to harness these data.

Data are expensive to collect, manage, and archive (Arzberger *et al.*, 2004; Tenopir *et al.*, 2011; Michener, 2015; Pisani *et al.*, 2016; Rockhold *et al.*, 2016). A failure to get these data from producers to users can lead to (i) lost opportunities to inform science, decision-making, and management (Vahedifard *et al.*, 2019) and (ii) result in costly replication of data collecting effort, both of which represent "data waste." There is much to be gained from finding new ways of reducing data waste to help manage the world's ocean.

A datashed framework for understanding data sharing and use

Like the various streams in a watershed, data flow from producer to user, often along a circuitous path, or data stream (Figure 2) creating a "datashed" that connects producers to users. In some cases, data never leave the instrument of collection, be it a hand-written notebook, sensor, or a smartphone (Thessen *et al.*, 2012; Hampton *et al.*, 2013; Ferguson *et al.*, 2014). Data may be held tightly by the researcher or company collecting them – to be converted into analysis, publications (Huang *et al.*, 2012), or even profit. Many datasets exist in a highly fragmented state (Chavan and Ingwersen, 2009; Reichman *et al.*, 2011). Some data are shared only locally, within a laboratory or a government agency, sites at which pools of data are generated. Smaller data pools may, over time, flow into larger data lakes, perhaps organized by institution, discipline, or project. As the number of data pools increases (Goodman *et al.*, 2014; Michener, 2015), individual data become more disconnected and harder to find.

The datashed that links producers to users is most effective for decision-making if data flow smoothly and can be tracked from producer to user, but numerous obstacles impede this flow (Reichman *et al.*, 2011; Fecher *et al.*, 2015; Tenopir *et al.*, 2015; Mbuagbaw *et al.*, 2017). Data dams, created by firewalls, paywalls, or layers of menus, can trap data or otherwise render them undiscoverable or inaccessible. Projects end and data stagnate (they cease to be updated or maintained) becoming trapped in swamps. Data may be delivered in a form that make them difficult to use beyond their original intent. As datasheds become longer, the ability of end users to communicate to producers and intermediary managers further upstream may decline. Even when one datashed flows without impediment, disciplinary mountains may separate datasheds (Kim and Zhang, 2015; Kim and Stanton, 2016), preventing the interdisciplinary integration of data that is necessary to study the whole-system thinking required to model ocean ecosystem change and to monitor the effectiveness of ocean policies (Gill *et al.*, 2017).

Barriers that need to be surmounted to promote data sharing and use

To effectively manage the global ocean and its living resources we need to remove data-dams, build links between datasheds, and

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

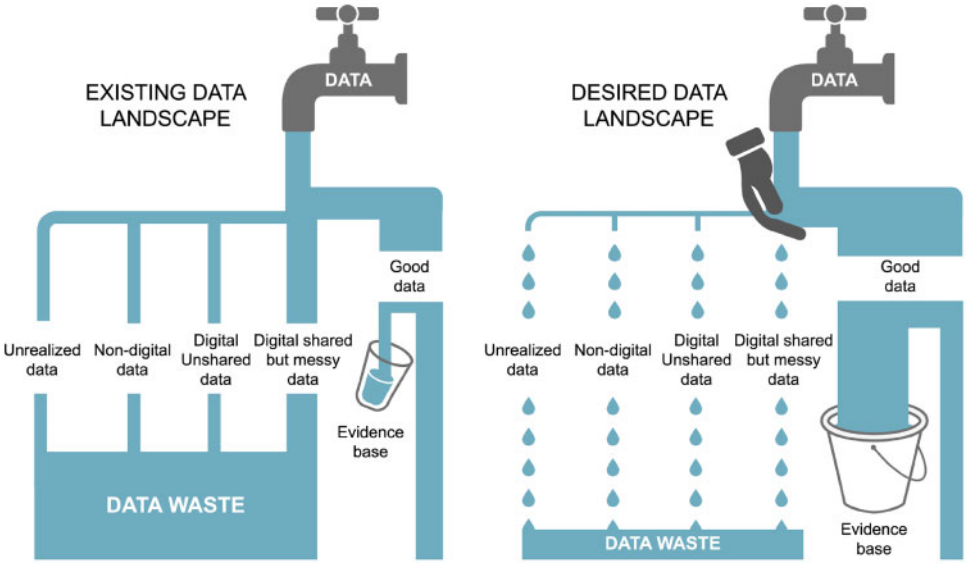

Figure 1. Current and desired states of data availability and sharing. The relative sizes of the water flows correspond to the amount of data we speculate falls within each data type.

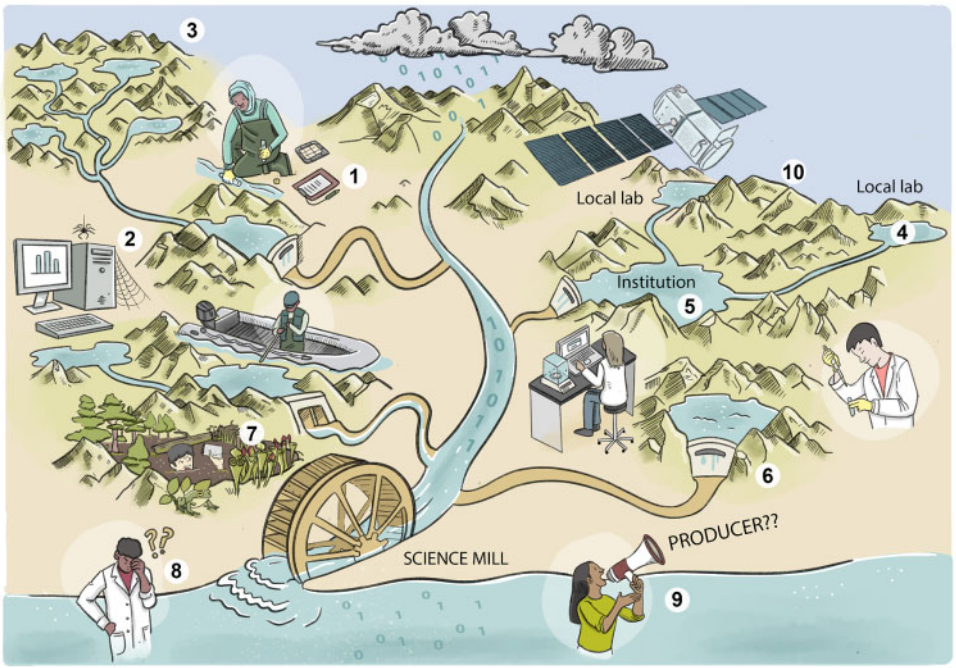

Figure 2. The "datashed" concept, in which data is generated by producers and, through processes of uploading and sharing, can become aggregated with other data "streams." (1) Data never leaves the instrument of collection; (2) data that are held tightly by the producer; (3) highly fragmented datasets; (4) local data sets that are shared only within the institutional body; (5) smaller data pools that may eventually pool together as part of larger projects involving multiple groups; (6) data dams, created by firewalls, paywalls, or layers of menus; (7) ended projects with their stagnated data; (8) data in unusable forms; (9) disassociation of data sets from their original producers; and (10) disciplinary mountains that separate datasheds.

find new and better ways to navigate the enormity of data, both in scale and scope. To do this, we identify three classes of challenges to data sharing (Figure 3): uploading, aggregating, and navigating (see Reichman *et al.*, 2011 for another typology).

Uploading

Uploading data to the cloud has become the first step needed to get data into a larger digital ecosystem where it can be shared. A number of obstacles inhibit data sharing including:

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

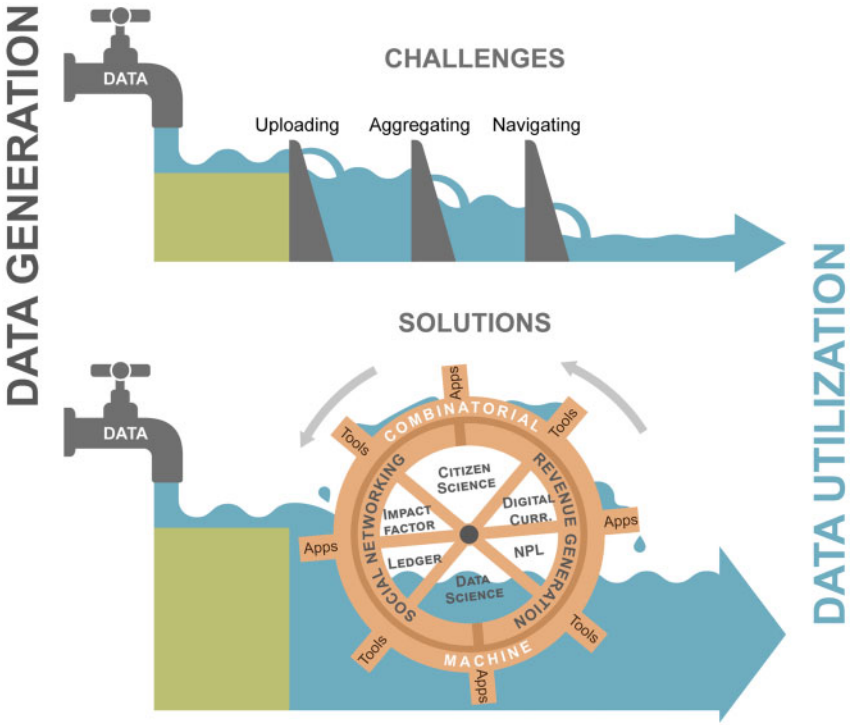

Figure 3. Three key barriers to data sharing and use include uploading, aggregating and navigating. Both technical and cultural solutions exist that could break down these barriers, and could be implemented within the context of a combinatorial machine (shown as a wheel) that itself provides a platform for discovery and access to data from many different tools and applications.

(1) authorship/ownership concerns (e.g. fear of being "scooped," pressure to publish, commercial proprietary interests, loss of control, and the fear that the data will be changed or critiqued; Costello, 2009; Reichman *et al.*, 2011; Baker, 2016);

(2) the time and effort required to find appropriate portals and upload data (Arzberger *et al.*, 2004; Tenopir *et al.*, 2011; Michener, 2015; Pisani *et al.*, 2016; Rockhold *et al.*, 2016; Wilkinson *et al.*, 2016; Park and Wolfram, 2017);

(3) completing often arduous metadata and data formatting requirements (Kim and Stanton, 2016); and

(4) a lack of incentives to share data (Costello, 2009; Reichman *et al.*, 2011).

In some cases, data are not digital (e.g. data that are in log books or published in hard copy) or the data owner does not know they have data that might be useful to others (e.g. photographic data on social media and photo sharing platforms) or even that they could be data producers or citizen scientists (e.g. smartphone owners in Indonesian coastal villages who could photograph fish they eat or record changes in mangroves or seagrass beds).

Aggregating

Even when data are uploaded to the cloud, they often remain segregated—by geography, by discipline (including non-marine disciplines), by quality, etc. (Kim and Stanton, 2016). The variety of data formats, units (Silvello, 2018), collection methods, and metadata standards can make the integration of data difficult. Links between data producers and users can be easily lost, making it difficult to inform users if there are changes or updates in the data, or for users to communicate to producers about potential problems with or improvements that could be made to the data. Better aggregation does not necessarily require that data are all homogenized, stored together, and treated as the same, but it does require a more centralized portal to access decentralized and disparate data.

Navigating

Navigating data becomes increasingly difficult as the volume and diversity of data grow. As new data and platforms are added (Michener, 2015; Wilkinson *et al.*, 2016; Park and Wolfram, 2017), it can be increasingly difficult to know what data are available (discoverability), to find the data you want (searchability), and to know which data are the most current or most accurate (Goodman *et al.*, 2014). When data of varying quality (Costello, 2009) are pooled, it may become more difficult to manage these quality differences transparently. The tools associated with each data portal may only meet the needs of a small set of users.

Data users and scientists need a more streamlined way of searching for and accessing data. Google's new Data Search Tool™ provides a single site for data searching, but does not provide access to the data sets. Without a better system for navigating data (aggregated virtually or otherwise), the transaction costs of searching and compiling the data required for highly interdisciplinary analysis over large geographic and temporal scales can be very high. A lack of a data integration framework or system has been blamed for the failure to respond to rapid-onset natural disasters like the Camp Fires in California (Vahedifard *et al.*, 2019).

Solutions

Both technical and cultural solutions exist (Table 1) that could break down the barriers to data flow (Figure 3). The key is to

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

Table 1. Cultural and technical solutions to break down barriers to data sharing and use.

	Impact factor	Cryptocurrency	Ledgers and block chain	NLP	Data science
Upload	×	×	×	×	
Aggregate			×	×	×
Navigate				×	×

combine technical and cultural approaches to create solutions and to draw from other sectors to learn how to do so. Here we review a short-list of potentially disruptive solutions that already exist, either at limited scale in the ocean data world or at more fully implemented scales in other fields.

Technical solutions

We define technical solutions as those that require limited or no behavioural change for data producers or users. In our framework, technical solutions solve existing problems by changing the technical nature of data or data sharing. These solutions could be even more powerful if combined with cultural solutions. Two types of technical solutions could be transformational to data sharing if applied at scale:

Ledger-based technologies can address issues of authorship, authenticity, traceability, immutability, and transparency, thus breaking down some barriers to sharing data. A ledger is a way of uniquely identifying data, associating that data with a particular author or owner, and possibly tracking it from source to final use (collectively known as provenance; Bell *et al.*, 2017; Hoy, 2017). Such ledgers can be differentiated by the degree to which they ensure trust, immutability, and provenance. The simplest ledger is the digital object identifier (DOI) already available for many types of data (e.g. https://datadryad.org/pages/repository). DOIs provide a minimum level of assurance regarding provenance—they are associated with the original data, but datasets that are transformed, or combined may have their own DOI, breaking the chain of provenance. Other types of ledgers, such as blockchain, could provide more information, traceability and security for data (Bell *et al.*, 2017; Extance, 2017; Bartling, 2018; Günther and Chirita, 2018; Pluto, 2018). Blockchains could record changes to data, both updates of original data and data transformation that occurs as part of secondary and tertiary analyses. In addition, they could be used to track data as it is combined into larger or more interdisciplinary data sets (e.g. Blockchain for science; Bartling, 2018). Permissioned blockchains could reflect the level of review, including peer review and thus embody information on quality (Rossum, 2017; Günther and Chirita, 2018; Pluto, 2018) leading to more transparency and accountability.

Automatic data translation and information extraction, such as natural language processing (NLP), automatic image analysis, and tools like Global Database of Events, Language and Tone (Leetaru and Schrodt, 2013), could leverage artificial intelligence to rapidly process large amounts of information, including widely varying human language input, images, and acoustic data. Such techniques could reduce costs and efforts associated with uploading and navigating data. NLP already is used by bibliographic software to read articles and papers and to automatically glean bibliographic data (Hull *et al.*, 2008). Similar algorithms could be used to automatically read methods and metadata descriptions and then populate metadata fields, thereby reducing the time required to upload data (Valdez *et al.*, 2016). NLP algorithms also could be used to convert data in white papers, journal papers, and logbooks into digital data (i.e. virtualization; Thessen *et al.*, 2012). Of course, accuracy of NLP-generated data and metadata would be greatly improved if verified by the data producers. The ledger-based methods described above could indicate whether such verification has occurred. Finally, automatic data discovery and NLP will be essential for any user interface to help data users find the data they need to answer the questions they have.

Cultural solutions

We define cultural solutions as those that result in significant changes to human behaviour. Cultural solutions might be those that engender trust and participation through largely social means as well as approaches that create new incentives for data sharing. Already, a command and control approach to require data sharing has been undertaken by many publications (e.g. requiring GenBank numbers for all DNA sequences) and granting agencies (e.g. EU's INSPIRE directive, the National Science Foundation), but these rules cannot guarantee that data are easily accessible, persistent, usable, or used. Nor do these rules necessarily encourage the sharing of private or commercially acquired data (One exception is GenBank, which has developed into an indispensible resource of DNA sequence data; Hampton *et al.*, 2013).

We need a new mindset along with new incentives and approaches that encourage voluntary data sharing. Here we review several potentially disruptive ideas.

An *ocean data impact factor* could incentivize data sharing, more primary data collection, and better data usability (Chavan and Ingwersen, 2009; Costello, 2009; Ferguson *et al.*, 2014; Kim and Stanton, 2016; Bierer *et al.*, 2017). For academics, quantifiable recognition is a powerful motivating factor. A more generic Data Citation Index℠ already exists for papers and datasets listed on the Web of Science. An impact factor or index for ocean data that more fully reflects how and how often data are used could encourage data to be shared in more than just the minimum standard required by journals and granting agencies. Data ledgers are necessary to quantify the intellectual impact of data. More elaborate ledgers (e.g. blockchains) could help account for the full life history of data and thus allow impact factors to reflect the full contribution of the original data set, including its use in transformed data and analyses (Günther and Chirita, 2018).

Digital community currencies could create new incentives for data sharing within the ocean data community. Community currencies are those, often unofficial, currencies that can only be exchanged in geographically designated areas (Naqvi and Southgate, 2013). Community currencies are often used to incentivize local spending, sometimes on specific products like local agriculture

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

produce or crafts. A digital community currency (DCC; Diniz *et al.*, 2018) could be created that rewards data providers based on the quantity, quality, or transparency of data provided or by responsiveness to questions from users. Unlike some favourability scores used by academic social networks like ResearchGate, a DCC could be exchanged for data services (e.g. archiving and storage, quality review and assurance, higher search result visibility, or better access to larger quantities of data). A DCC would likely be most influential in spurring data sharing by data providers who have large amounts of data to share and for those entities that may have both a willingness and capacity to pay to use data (e.g. private companies that provide services like ship routing or government agencies that need to manage natural disasters).

Social networks have proved highly disruptive in the sharing of images, knowledge, and commentary for non-scientific sectors (e.g. Facebook, Instagram) and more recently for sharing publications and knowledge (e.g. ResearchGate, Academia.com; Noorden, 2014). Furthermore, social media has been shown to build trust between consumers and brands in the marketing space (Schmidt and Iyer, 2015). Because trust is a key element in the decision to share data (Reichman *et al.*, 2011), efforts to connect researchers with other researchers and to end-users could increase trust and data sharing. Adding social networking to online data platforms could help share information about data quality, usefulness, updates, and changes and to provide feedback on how data are used and collected (e.g. through sentiment analysis). If closely integrated with data platforms, social networking could not only would contribute to better transparency, it could provide a way of letting "the crowd" gauge quality, and could help build trust between users and potential data providers.

Social networks, combined with data platforms, also could facilitate connections and exchanges that in turn could spur ideation and knowledge exchange while providing an avenue for users and other researchers to announce data needs or to signal to others they may be available to help in data collection. Just as some ridesharing applications leverage social networking to help riders find drivers, social networks could also help researchers coordinate field research by letting others know when and where they are going into the field or by communicating through the network about particular needs for help in collecting data.

While social networks are not without flaws (e.g. fake reviews; Kumar *et al.*, 2018), many options are being developed to address these issues and to enhance the community building opportunities for their use.

Conclusion: could a data combinatorial machine disrupt ocean data sharing?

Data sharing platforms for the ocean are emerging rapidly—each with a different intended audience. Some platforms focus solely on the data while others focus on the social and cultural dimensions of researchers and users. Some are built around a particular analytical or visualization tool or interface. We applaud these efforts and propose an over-arching umbrella, an *ocean data combinatorial machine* (*ODCM*), that could serve to "virtually" create a data and social foundation that would bring together data, researchers, and users and would also build a foundation upon which analytical tools could be built.

Combinatorial machines (CMs) are technology platforms that can combine aggregating and navigating technologies and social networks. Amazon.com, Alibaba, TripAdvisor, and other commercial CMs have solved many problems similar to those faced by the ocean data sector, but applied to consumer products, markets, and travel. For instance, Amazon is an aggregation centre that uses real warehouses and advanced distribution centres to find the right combination of "storage" and on-demand access to consumer items. Only selected items are kept on-hand. Amazon also increasingly provides more and more details about the origin, composition, and technical details of products—essentially the metadata of consumer goods. The built-in social networking at Amazon allows consumers to: (i) rate the quality of products; (ii) discuss experiences using the product; and (iii) in many cases interact directly with vendors and producers.

Drawing on lessons learned from commercial CMs, we propose a new type of ODCM that would create a more centralized way of bringing data, data producers, and data users together. Instead of exchanging packages of consumer goods or travel advice, an ODCM would facilitate the exchange of packets of data. Like Amazon.com, such an ODCM would not need to be a physical home to all data (and perhaps no data). Application programming interfaces (APIs), software architectural design approaches (e.g. representational state transfer), web crawling, machine learning, and other technologies could be used to virtually pull data from individual producers, passive data collectors, open-data pools, and permissioned-data centres into a single platform into the ODCM as needed. An ODCM could support NLP, case-based reasoning, and affinity analysis to improve the human-data interface, making searching easier, and recommending potentially relevant data, researchers, and research projects. An ODCM could become a convenient hub for data-related social networks—allowing users to comment directly on data quality, how data were used, direct links to where data has contributed to publications, updates, and other issues in the same way one can review a commercial product. Like Amazon, an ODCM would not require homogeneity in data quality or standards, but would benefit from transparency about metadata, data origin, data quality, and data provenance.

An ODCM could provide a platform upon which third-party analytical applications and tools, both freely available and commercial, could be built that could facilitate the use of ocean data by a wider variety of users. For instance, applications could be added to the ODCM that translate different data standards or units to common formats for selected types of data. Data self-assessment tools like the ARDC Fair Data Tool (https://www.ands-nectar-rds.org.au/fair-tool) could be linked directly to the ODCM to help data producers and providers (and the general public) determine the degree to which individual datasets meet FAIR standards (e.g. findable, accessible, interoperable, and reusable; Wilkinson *et al.*, 2016). Mapping applications, habitat suitability modelling tools (e.g. Octopus, https://octopus.zoo.ox.ac.uk/beta), and other apps could be built as needed by app-developers.

Each of the tools proposed above already exists, but not always specifically for ocean data. No data platform yet exists that combines all of the above approaches for ocean data. The next step is to determine how to build upon what we already know from other sectors and how to do so in a way that creates a financially

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

sustainable ocean data ecosystem (Warren, 2016). While public funding of data collection and many data lakes and portals will always be a part of the ocean data seascape, we must do more to think out-of-the-box to find ways of making an ODCM financially viable. While many types of data must remain open, proprietary data could be exchanged within the ODCM through standard commercial (for hire) agreements, made possible by the contracts implicit in blockchains. Like commercial CMs, the ODCM could be made more financially viable through the use of revenue generating targeted advertisements (e.g. by journals and scientific instrument companies, such as used to support ResearchGate) and faster, larger downloading options could be made available for a fee. Finding new business models for data platforms will be key to their success and sustainability. While the ODCM should not present financial barriers to access open data, it could charge for using enhanced access, navigation, and other services provided by the ODCM and its community (similar to the model employed by commercial weather, wind, and wave forecasting services).

An ODCM should be designed so that it does not represent an added layer of complexity to an already highly complex universe of ocean data. Instead, the ODCM as we envision it would help data producers and users better manage the increasingly complex and messy world of ocean data and to fully benefit from the information and understanding that could come from harnessing that complexity. Combined with the rapid advances in data science and visualization, the ODCM could provide better access to the raw information that a new breed of ocean scientists and planners need to measure, model, and manage the ocean. Finally, an ODCM could provide the critical mass needed to create a new, global community of researchers, citizen scientists, data producers and users, government analysts, and others that could, itself, transform the way we source and share ocean data.

The future of ocean health, and the planet's health, depends on our ability to coordinate effective action to resolve the drivers of environmental degradation. Improved access to and use of data will greatly enhance our ability to plan, implement and monitor the impacts of policy and management. Advances in data science and social networking provide great hope and opportunity that we can revolutionize the way we collect and use ocean data, but only if we look to collaborate outside the ocean sector to make the best use of what others have achieved.

Author contributions

LHP initiated the research and led the development of the conceptual design, and drafted the first version of the paper. SOG, DBK, and CN collected background material, contributed text to all sections of the paper. AS provided input on data portals and conservation science use of data and participated in the research retreat. HB, E, OHG, EK, LL, and the IBM Team contributed to the conceptual design and improvements to the manuscript. Figures were drawn by Vanessa González-Ortiz (http://vgonzalezortiz.com) who retains all copyrights.

Supplementary data

Supplementary material is available at the *ICESJMS* online version of the manuscript.

Acknowledgements

We would like to thank Emma Camp, Sophie Dove, and Catherine Lovelock for providing input regarding marine conservation science. Richard Leck and John Tanzer provided input into the use of data for marine conservation planning. IBM's Corporate Services Corp provided the financial resources that enabled the participation of the IBM CSC AUS1 team. Australian Business Ventures organized the logistics of the month-long research retreat that led to this paper. The Global Change Institute at the University of Queensland hosted this research project.

Funding

These ideas came out of a disruptive, month-long research retreat made possible by the IBM Corporate Service Corps (CSC), WWF- Global Science, WWF-US, and the Global Change Institute (CGI) at the University of Queensland. All of these organizations contributed in kind (e.g. space and research time).

Conflict of interest statement

No authors have any competing interests. WWF-Norway is a partner of REV Oceans which has proposed an Ocean Data Platform (ODP). Some of the ideas in this paper have been shared with the ODP.

References

Arzberger, P., Schroeder, P., Beaulieu, A., Bowker, G., Casey, K., Laaksonen, L., and Moorman, D. 2004. Promoting access to public research data for scientific, economic, and social development. Data Science Journal, 3: 135–152.

Baker, M. 2016. Is there a reproducibility crisis? Nature, 533: 452–454.

Bartling, S. 2018. Blockchain for Science. https://docs.google.com/document/d/1Uhjb4K69l0bSx7UXYUStV_rjuPC7VGo0ERa-7xEsr58/edit (last accessed 22 November 2018).

Bell, J., Latoza, T. D., Baldmitsi, F., and Stavrou, A. 2017. Advancing open science with version control and blockchains. Proceedings—2017 IEEE/ACM 12th International Workshop on Software Engineering for Science. SE4Science, 2017: 13–14.

Bierer, B. E., Crosas, M., and Pierce, H. H. 2017. Data authorship as an incentive to data sharing. The New England Journal of Medicine, 373: 567–571.

Chavan, V. S., and Ingwersen, P. 2009. Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community. BMC Bioinformatics, 10: 1–11.

Costello, M. J. 2009. Motivating online publication of data. BioScience, 59: 418–427.

Costello, M. J., Coll, M., Danovaro, R., Halpin, P., Ojaveer, H., and Miloslavich, P. 2010. A census of marine biodiversity knowledge, resources, and future challenges. PLoS One, 5: e12110–e12126.

Diniz, E. H., Siqueira, E. S., and van Heck, E. 2018. Taxonomy of digital community currency platforms. Information Technology for Development, 25: 69–91.

Extance, A. 2017. Blockchain moves to science. Nature, 552: 301–302.

FAO. 2018. The State of World Fisheries and Aquaculture 2018—Meeting the Sustainable Development Goals. FAO, Rome. 2 pp. www.fao.org/3/i9540en/I9540EN.pdf (last accessed 7 January 2019).

Fecher, B., Friesike, S., and Hebing, M. 2015. What drives academic data sharing? PLoS One, 10: 1–25.

Ferguson, A. R., Nielson, J. L., Cragin, M. H., Bandrowski, A. E., and Martone, M. E. 2014. Big data from small data: data-sharing in the 'long tail' of neuroscience. Nature Neuroscience, 17: 1442–1447.

Fisher, B., Balmford, A., Ferraro, P. J., Glew, L., Mascia, M., Naidoo, R., and Ricketts, T. H. 2014. Moving Rio forward and avoiding 10

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

more years with little evidence for effective conservation policy. Conservation Biology, 28: 880–882.

Gill, D. A., Mascia, M. B., Ahmadia, G. N., Glew, L., Lester, S. E., Barnes, M., Craigie, I. *et al.* 2017. Capacity shortfalls hinder the performance of marine protected areas globally. Nature, 543: 665–669. http://www.nature.com/articles/nature21708 (last accessed 19 November 2018).

Goodman, A., Pepe, A., Blocker, A. W., Borgman, C. L., Cranmer, K., Crosas, M., Stefano, R. D. *et al.* 2014. Ten simple rules for the care and feeding of scientific data. PLoS Computational Biology, 10: e1003542.

Günther, V., and Chirita, A. 2018. "Scienceroot" Whitepaper. https://www.scienceroot.com/ (last accessed 22 November 2018).

Hampton, S. E., Strasser, C. A., Tewksbury, J. J., Gram, W. K., Budden, A. E., Batcheller, A. L., Duke, C. S. *et al.* 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment, 11: 156–162.

Haward, M. 2018. Plastic pollution of the world's seas and oceans as a contemporary challenge in ocean governance. Nature Communications, 9: 667.

Hoegh-Guldberg, O., Jacob, D., Taylor, M., Bindi, M., Brown, S., Camilloni, I., Diedhiou, A. *et al.* 2018. Impacts of 1.5¤C Global Warming on Natural and Human Systems. In: Global Warming of 1.5°C. An IPCC Special Report on the impacts of global warming of 1.5°C above pre-industrial levels and related global greenhouse gas emission pathways, in the context of strengthening the global response to the threat of climate change, sustainable development, and efforts to eradicate poverty. http://www.ipcc.ch/report/sr15 (last accessed 15 November 2018). In press.

Hoegh-Guldberg, O., Jacob, D., and Taylor, M. 2018. Impacts of 1.5°C global warming on natural and human systems. *In* IPCC SR 1.5. 243 pp.

Hoy, M. B. 2017. An introduction to the blockchain and its implications for libraries and medicine. Medical Reference Services Quarterly, 36: 273–279.

Huang, X., Hawkins, B. A., Lei, F., Miller, G. L., Favret, C., Zhang, R., and Qiao, G. 2012. Willing or unwilling to share primary biodiversity data: results and implications of an international survey. Conservation Letters, 5: 399–406.

Hull, D., Pettifer, S. R., and Kell, D. B. 2008. Defrosting the digital library: bibliographic tools for the next generation web. PLoS Computational Biology, 4: e1000204.

IOC-UNESCO. 2017. Global Ocean Science. UNESCO Publishing, Paris.

IUCN. 2018. The IUCN Red List of Threatened Species. Version 2018-2. http://www.iucnredlist.org (last accessed 16 November 2018).

Jamieson, A. J., Malkocs, T., Piertney, S. B., Fujii, T., and Zhang, Z. 2017. Bioaccumulation of persistent organic pollutants in the deepest ocean fauna. Nature Ecology and Evolution, 1: 24–27.

Kim, Y., and Zhang, P. 2015. Understanding data sharing behaviors of STEM researchers: the roles of attitudes, norms, and data repositories. Library and Information Science Research, 37: 189–200.

Kim, Y., and Stanton, J. M. 2016. Institutional and individual factors affecting scientists' data-sharing behaviors: a multilevel analysis. Journal of the Association for Information Science and Technology, 67: 776–799.

Kumar, N., Venugopal, D., Qiu, L., and Kumar, S. 2018. Detecting review manipulation on online platforms with hierarchical supervised learning. Journal of Management Information Systems, 35: 350–380.

Leetaru, K., and Schrodt, P. A. 2013. Gdelt. Eventdata.Psu.Edu: 1979–2012. http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf (last accessed 22 November 2018).

Mbuagbaw, L., Foster, G., Cheng, J., and Thabane, L. 2017. Challenges to complete and useful data sharing. Trials, 18: 17–19.

Michener, W. K. 2015. Ecological data sharing. Ecological Informatics, 29: 33–44.

Munthe, C., and Jensen, F. 20 November 2018. Sperm Whale Washed Up in Indonesia had Plastic Bottles, Bags in Stomach. Reuters, Jakarta. https://www.reuters.com/article/us-indonesia-whale/sperm-whale-washed-up-in-indonesia-had-plastic-bottles-bags-in-stomach-idUSKCN1NP11F (last accessed 10 January 2019).

Naqvi, M., and Southgate, J. 2013. Banknotes, local currencies and central bank objectives. Bank of England - Quarterly Bulletin, Q4: 317–325.

Noorden, R. V. 2014. Scientists and the social network. Nature, 512: 126.

Park, H., and Wolfram, D. 2017. An examination of research data sharing and re-use: implications for data citation practice. Scientometrics, 111: 443–461.

Parsons, E. C. M. 2016. Why IUCN should replace "data deficient" conservation status with a precautionary "assume threatened" status—a cetacean case study. Frontiers in Marine Science, 3: 2015–2017.

Pisani, E., Aaby, P., Breugelmans, J. G., Carr, D., Groves, T., Helinski, M., Kamuya, D. *et al.* 2016. Beyond open data: realising the health benefits of sharing data. BMJ (Clinical Research Edition), 355: i5295. doi: 10.1136/bmj.i5295.

Pluto. 2018. Pluto – Breaking down the barriers in academia. White Paper. 32 pp. https://assets.pluto.network/Pluto_white_paper_v04_180719_1355_BSH.pdf (last accessed 22 November 2018).

Poloczanska, E. S., Burrows, M. T., Brown, C.J.G., Molinos, J., Halpern, B. S., Hoegh-Guldberg, O., Kappel, C. V. *et al.* 2016. Responses of marine organisms to climate change across oceans. Frontiers in Marine Science, 3: 1–21. http://journal.frontiersin.org/Article/10.3389/fmars.2016.00062/abstract (last accessed 15 December 2018).

Reichman, O. J., Jones, M. B., and Schildhauer, M. P. 2011. Challenges and opportunities of open data in ecology. Science, 331: 703–705.

Rockhold, F., Nisen, P., and Freeman, A. 2016. Data sharing at a crossroads. New England Journal of Medicine, 375: 1112–1115.

Rossum, J. v. 2017. Blockchain for research. Philologus, 71: 566–567.

Sandwell, D., Gille, S., Orcutt, J., and Smith, W. 2003. Bathymetry from space is now possible. Eos, 84: 37–44.

Schmidt, K. N., and Iyer, M. K. S. 2015. Online behaviour of social media participants' and perception of trust, comparing social media brand community groups and associated organized marketing strategies. Procedia - Social and Behavioral Sciences, 177: 432–439.

Science 20. 2019. S20 Japan 2019 Science 20 Threats to Coastal and Marine Ecosystems, and Conservation of the Ocean Environment – With Special Attention to Climate Change and Marine Plastic Waste www.scj.go.jp/ja/info/kohyo/pdf/kohyo-24-s20jp2019-1.pdf (last accessed 20 February 2019).

Silvello, G. 2018. Theory and practice of data citation. Journal of the Association for Information Science and Technology, 69: 6–20.

Sutherland, W. J., Pullin, A. S., Dolman, P. M., and Knight, T. M. 2004. The need for evidence-based conservation. Trends in Ecology and Evolution, 19: 305–308. https://www.sciencedirect.com/science/article/pii/S0169534704000734 (last accessed 19 November 2018).

Tenopir, C., Allard, S., Douglass, K., Aydinoglu, A. U., Wu, L., Read, E., Manoff, M. *et al.* 2011. Data sharing by scientists: practices and perceptions. PLoS One, 6: 1–21.

Tenopir, C., Dalton, E. D., Allard, S., Frame, M., Pjesivac, I., Birch, B., Pollock, D. *et al.* 2015. changes in data sharing and data reuse practices and perceptions among scientists worldwide. PLoS One, 10: e0134826.

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020

Thessen, A. E., Cui, H., and Mozzherin, D. 2012. Applications of natural language processing in biodiversity science. Advances in Bioinformatics, 2012: 391574.

Vahedifard, F., Ermagun, A., Mortezaei, K., and AghaKouchak, A. 2019. Integrated data could augment resilience. Science, 363: 134.

Valdez, J., Rueschman, M., Kim, M., Redline, S., and Sahoo, S. S. 2016. An ontology-enabled natural language processing pipeline for provenance metadata extraction from biomedical text. *In* OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10033. LNCS, Springer, Cham, pp. 699–708.

Visbeck, M. 2018. Ocean science research is key for a sustainable future. Nature Communications, 9: 1–4.

Warren, E. 2016. Strengthening research through data sharing. New England Journal of Medicine, 375: 401–403.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J. J., Appleton, G., Axton, M., Baak, A., Blomberg, N. *et al.* 2016. The FAIR guiding principles for scientific data management and stewardship. Scientific Data, 3: 160018. doi: 10.1038/sdata.2016.18.

WWF. 2018. Living Planet Report - 2018 Aiming Higher. World Wide Fund for Nature, Gland, Switzerland. http://www.panda.org/about_our_earth/all_publications/living_planet_report/ (last accessed 25 October 2018).

Handling editor: Robert Blasiak

Downloaded from https://academic.oup.com/icesjms/article/76/6/1415/5480138 by guest on 25 November 2020