Please cite as:
Durden, JM, Luo, JY, Alexander, H, Flanagan, AM & L Grossmann, 2017. Integrating "Big Data" into Aquatic Ecoogy: Challenges and Opportunities. *Limnology and Oceanography Bulletin*, doi: 10.1002/lob.10213

Integrating "Big Data" into Aquatic Ecology: Challenges and Opportunities

Jennifer M. Durden[1,2,3,*], Jessica Y. Luo[4], Harriet Alexander[5], Alison M. Flanagan[6], Lars Grossmann[7]

[1]Norwegian University of Science and Technology, Trondheim, Norway

[2]National Oceanography Centre, Southampton, UK

[3]Ocean and Earth Science, University of Southampton, National Oceanography Centre, Southampton, UK

[4]Climate & Global Dynamics Lab, National Center for Atmospheric Research, Boulder, Colorado, USA

[5]Department of Population Health and Reproduction, University of California, Davis, USA

[6]School of Marine & Atmospheric Sciences, Stony Brook University, Stony Brook, New York, USA

[7]Biodiversity Department and Centre for Water and Environmental Research, University of Duisburg-Essen, Essen, Germany

*corresponding author: jennifer.durden@ntnu.no

Abstract

Got 'Big Data'? Not sure how best to use it? Big Data is becoming an important facet of aquatic ecology, and researchers must learn to harness it to reap the rewards of using it. The benefits of using Big Data are many, and include advancements in scientific understanding at larger scales and higher resolution, applications to improving environmental management and policy, and public engagement. We aim to demystify the use of Big Data for individual scientists, and provide some food for thought for the aquatic ecology community on how to develop this sphere. To achieve this, we highlight six key challenges: 1) how to recognize if you have Big Data, 2) handling Big Data, 3) issues with classical analytical techniques, 4) verification of Big Data, 5) considerations for data sharing, and 6) community development of knowledge infrastructures. We then present approaches and tools which have been successfully applied to these challenges in aquatic ecology and other scientific fields.

Introduction

The words "Big Data" elicit a variety of responses from aquatic ecologists, from glee at the possibilities for ecological understanding at a wide range of spatiotemporal scales, to dread at the daunting task of data management. Big Data spans scales from climate science to molecular biology, and has facilitated discoveries from 'macrosystems ecology' at regional scales (Soranno and Schimel 2014) to the molecular underpinnings of competition between plankton (Alexander et al. 2015). The integration of Big Data into aquatic ecology has the potential to change its foundational theories, as it has in other areas of research (Kitchin 2014). Other benefits of using Big Data include advancements in scientific understanding at larger scales and higher resolution, multi- and cross-scale analysis of patterns (Soranno and Schimel 2014), applications to improving environmental management and policy, and public engagement (e.g. through crowd sourcing, Matabos et al. 2017).

Despite these benefits, barriers to using Big Data often prevent its optimal use. A primary issue is that data management problems for the user are real (Borgman 2015a; Boyle 2013; Hampton et al. 2013). The challenge of amassing, storing, analysing, and preserving large volumes of diverse data types is not a minor one. For example, collecting Big Data using new technology may present challenges with data velocity, while compiling disparate datasets from others to form Big Data presents challenges of standardization (Soranno et al. 2015a). Cultural barriers to data sharing exist, and developing infrastructures to facilitate sharing is a related challenge (Borgman 2015a; Borgman et al. 2015). Furthermore, traditional approaches to analysis may not be appropriate, and new methods must be developed (e.g. Fei et al. 2016; Hooten and Hobbs 2015).

Optimists should take heart – Big Data has been wrangled in the past, producing a step change in the field of biology with the production of the Linnaeus classification system (Müller-Wille and Charmantier 2012), and in the management and distribution of data generated by the Large Hadron Collider at CERN (Allcock et al. 2002). So, how can we best move forward, as individual users and as a research community?

Big Data is used in other scientific disciplines, such as astrophysics, economics, and material sciences, and we can learn from their approaches. Aquatic scientists have an opportunity to bridge gaps between disciplines with similar techniques, such as atmospheric or terrestrial sciences. However, the application of Big Data presents some particular challenges, which are common but not exclusive to aquatic ecology:

- Aquatic ecology involves diverse types of data (high data "variety"). These include data associated with many methods of measurement (e.g. instrument-specific, continuous / discrete / qualitative).
- These data exist at a wide range of scales, in terms of both resolution and extent.
- Ecological datasets also include environmental variables, such as physical, chemical, and geological data on the physical environment or habitat.
- These data are increasing in multiple dimensions (horizontal, vertical, time), and the resolution in different dimensions may vary.

Here, we highlight six key challenges in using Big Data in aquatic ecology and provide practical solutions for the user and the community.

1 Recognizing Big Data

How do you know when you have Big Data? It is a relative concept; for example, in deep-sea ecology, hundreds of thousands of seabed photographs constitute Big Data (e.g. Morris et al. 2014), while in 'omics research, hundreds of millions of data points are involved (e.g. Alexander et al. 2015). Thus, we recommend that aquatic scientists consider their data in the context of other similar aquatic ecology data.

'Big science' groups, such as large collaborations that share instruments and infrastructure across institutions, are more obvious generators of Big Data than researchers that comprise small groups using local resources (Borgman 2015a). Ecology has previously been described as 'little science', in which 'artisanal' data are acquired using locally-made tools and tailored methods, with little replication due to ecosystem dynamics (Bowen and Roth 2007). However, developments in technology are increasing automated data collection over artisanal data, with accompanying increases in the velocity and volume of data collected (see Fig. 1). Examples of aquatic ecology methods that are moving data collection from artisanal to more automated include: underwater photography using autonomous vehicles with some automated annotation for benthic community estimation (Durden et al. 2016; Morris

et al. 2014; Schoening et al. 2012); a sedimentation event sensor used to quantify organic matter deposition in lieu of sediment traps (McGill et al. 2016); optical and acoustic sensors for identifying seabed habitats as opposed to interpolating physical point samples (Costa et al. 2009; Flanagan 2016); and underway continuous flow cytometric samplers (Swalwell et al. 2011) and the Scripps Plankton Camera (spc.ucsd.edu). Large interdisciplinary sampling efforts in aquatic ecology, such as ocean observatories (Favali et al. 2015), components of the US Large Ecological Time Series (Hobbie et al. 2003) and oceanic expeditions such as Tara Oceans, (Hobbie et al. 2003), are also contributing to Big Data generation, and large-scale comparisons. Small datasets can also be aggregated or accumulated to form Big Data, for example from long-term time series (e.g. Hampton et al. 2008; Smith et al. 2013), or the aggregation of of regional monitoring datasets into sub-continental scale water quality databases (Soranno et al. 2015a).

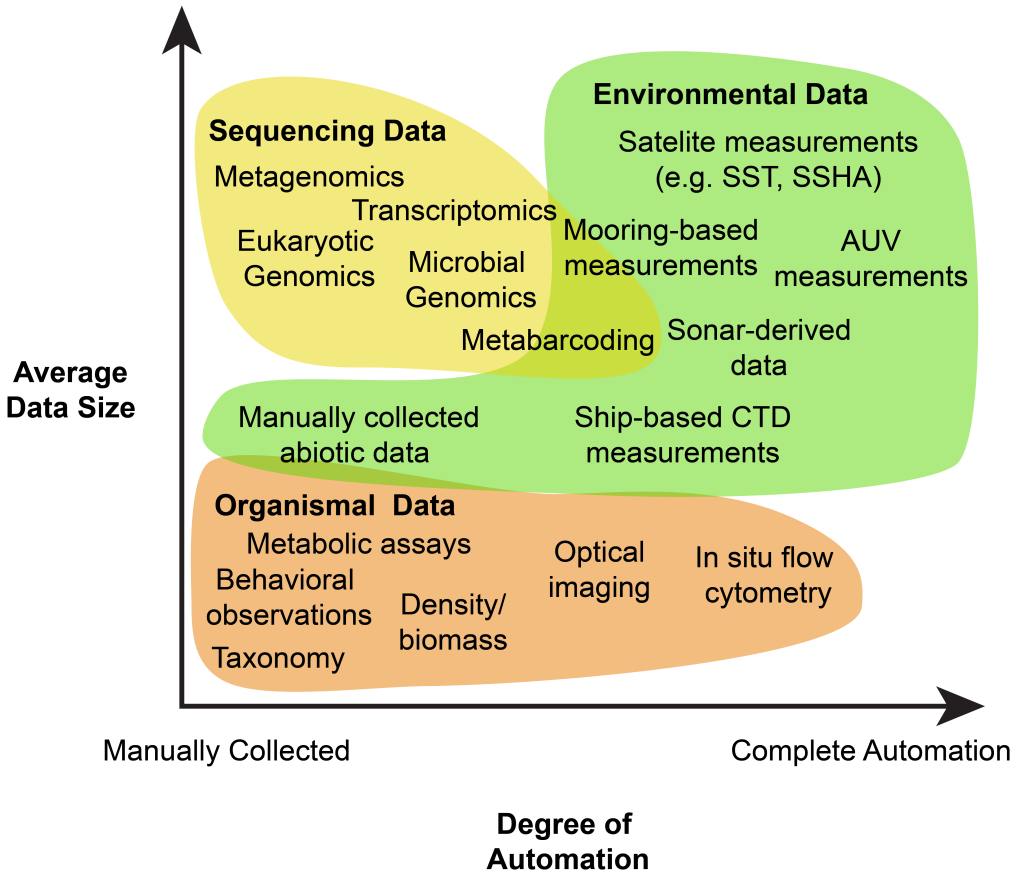

Figure 1. Examples of data used by aquatic ecologists, defined by their method of collection. Automatically-captured data is more likely to become Big Data than artisanally-captured data.

2 Data handling

To begin analyzing and interpreting Big Data, scientists must first overcome the hurdles associated with the handling, management, and manipulation of these large datasets (Mattmann 2013). While gathering data used to be the limiting step in data analysis, it has now been supplanted by the storage, transfer, and processing of these large data types. Computing is being recognized as a keystone step in the scientific process. This recognition highlights the need for computational savvy

aquatic scientists, better access to computational platforms, improved software tools and development, and database/archival management. We suggest that the following areas be considered:

1. *Computational preparedness:* Now, more than ever, aquatic scientists must be computationally literate. Soranno et al. (2015a) noted that this must go beyond training in quantitative analysis, and include skills to engage in data-intensive science, including interacting with large databases. An increased emphasis should be placed on computational and statistical training not only at the graduate level, but at the undergraduate level, weaving computational and statistical components into aquatic science curricula. Emphasis should be placed on learning scientific computing (e.g. R, Python, C++) at earlier stages in a research career. Additionally, active scientists and graduate students should be encouraged and supported in the pursuit of ongoing computational training through programs such as Software Carpentry and Data Carpentry (Wilson et al. 2014). Collaboration between ecologists and computer scientists and engineers should also be encouraged to tackle complex projects.

2. *Scientific software and pipeline creation*: Emphasis should be placed upon the computational methods used in the analysis of aquatic datasets. As data analysis becomes increasingly complex, integrated processing pipelines (a set of scripts for automating data analysis) become essential for reproducibility and efficiency (Goodman et al. 2014). Documentation of such pipelines and processes should be as transparent as laboratory-based manipulations. Best practices, such as the use of version control (e.g., Git, SVN; Perez-Riverol et al. 2016) and documentation of the design and purpose of code (Wilson et al. 2014), should become common practice and required for publication. Tools, such as protocols.io, promise user-friendly means of generating and publishing such workflows (Teytelman et al. 2016).

3. *Computational platforms:* Access to high-performance computational (HPC) resources is central to the analysis of Big Data. Different types of data, analytical software, and algorithms have different limitations. For example, bioinformatics analysis is often memory-bound while image analysis or modeling is more typically input/output or central processing unit-bound. As such, they are optimized on different computational architectures. While access to a well-managed high-performance computing system may be ideal in many circumstances, access is often institution-specific and may be limited in size or scope. Cloud computation, such as Amazon Web Service or Google Cloud Platform, provides a viable, if more expensive, alternative to high performance computing that is often more flexible than other options.

4. *Data transfer:* Big Data is now generated during field campaigns (e.g. digital photographs/video, acoustic data, automated sensor data), presenting a challenge in transferring this data from the field storage platform to facilities with HPC resources for processing analysis. While wireless data transfers from sensors are increasingly used in shallow water or moored instrument applications, physical hard drives or RAID disks are still used in transferring data acquired from field campaigns undertaken far from land. Transferring data from a single hard drive to an HPC array becomes resource intensive as data scales up in volume, particularly if storage types and transfer modes are not completely compatible. No single solution exists in this case, but explicit consideration of this process should be included in data management plans. Important factors to consider up-front are the compatibility of operating systems and hard drives, and the often-limiting machine input/output performance. Consulting with a HPC / data management expert prior to initiating a large field campaign may assist in the development of unique solutions to an often overlooked, but potentially limiting, problem in Big Data.

Please cite as:
Durden, JM, Luo, JY, Alexander, H, Flanagan, AM & L Grossmann, 2017. Integrating "Big Data" into Aquatic Ecoogy: Challenges and Opportunities. *Limnology and Oceanography Bulletin*, doi: 10.1002/lob.10213

3 Analytical Techniques

Aquatic ecologists who have been catapulted into the 'Big Data era' cannot rely on testing hypotheses with classical frequentist statistics, such as analysis of variance, t-tests, and linear regression. Big Data presents statistical challenges including enormous sample sizes, spurious correlation among explanatory variables, zero-inflated datasets, non-normality, and spatial/temporal autocorrelation (e.g. Dray et al. 2012; Fan et al. 2014; Legendre 1993). These statistical challenges mean that classical frequentist statistics are inadequate for Big Data ecology. Fortunately, there are solutions. Here we highlight some aspects of data analysis that are particularly relevant to Big Data.

1. *Significance:* One of the primary issues with large sample sizes is that many standard statistical test can be declared "significant" if the sample size is large enough (e.g., Sullivan and Feinn 2012). This happens because even minute differences from 0 (the null hypothesis) can be detected, leading to spurious yet statistically significant results. Ultimately, this leads to an increased risk of Type I Error (a 'false positive'), a big problem in Big Data ecology. Solutions involve moving beyond the use of classical frequentist statistics (p-value testing) to more advanced analyses (see point 3 below).

2. *Normality:* Ecological data tend to be non-normally distributed, and require transformation (e.g. logarithm or square root) for datasets to conform to assumptions of normality (Sokal and Rohlf 1995). This is exacerbated with Big Data, as datasets with extensive spatiotemporal coverage also typically contain large numbers of zeros; therefore, standard transformations are typically insufficient. Zero-inflated datasets can be transformed prior to multivariate analysis (Legendre and Gallagher 2001), or analyzed using hurdle models with zero-inflated Poisson or negative binomial distributions (Potts and Elith 2006; Seiler et al. 2012; Ver Hoef and Jansen 2007).

3. *Independence and autocorrelation:* Independence is a key criterion for many statistics, and violating this assumption can give misleading results. For a given location in time, it is unlikely that any two biotic or environmental measurements are truly "independent", and aquatic ecologists should be wary of spurious correlations between variables. Spatiotemporal autocorrelation (Legendre 1993) is another major statistical challenge with large ecological datasets that must be addressed when analysing Big Data.

Multivariate ordination techniques (e.g., canonical correspondence, redundancy analysis) and spatial statistics (e.g., multiscale ordination; Wagner 2003; Wagner 2004) provide new ways of analyzing large datasets that allow us to address challenges such as enormous sample sizes, spurious correlation among explanatory variables, zero-inflation, non-normality, and autocorrelation. Mixed effect models (generalized linear and general additive mixed models) may circumvent autocorrelation by incorporating random effects, including those that are common in hierarchical nested sampling designs. Explicitly incorporating the analysis of autocorrelation (in residuals) using multiscale ordination has been suggested as an opportunity for directly incorporating spatial structure of biotic-environmental relationships as a proxy for abiotic factors that are difficult or essentially impossible to measure (Dray et al. 2012). Multiple testing corrections, such as Bonferroni (Bonferroni 1935) and Benjamini-Hochberg (Benjamini and Hochberg 1995), are used to address statistical issues in multiple comparisons of Big Data, in addition to resampling methods, such as bootstrapping and Monte-Carlo simulations.

4. *Bayesian approaches:* Analysis of ecological data using Bayesian approaches, rather than using frequentist statistics, has increased substantially with the advent of Big Data. Since these approaches assess the probability of a

hypothesis being true, they are advantageous in quantifying uncertainty, in gaining understanding from 'noisy' observational data, and in assessing cross-scale interactions (Levy et al. 2014). However, Bayesian inference should not be interpreted similarly to inductive inference (Gelman and Shalizi 2013). Hooten and Hobbs (2015) reviewed Bayesian approaches to model selection particularly for ecologists, including the advantages and disadvantages of each. Methods for using Bayesian approaches that may be of interest to aquatic ecologists include applications to large datasets of animal movement in state-space modelling (Jonsen et al. 2003), use in predictive habitat distribution modelling (Guisan and Zimmermann 2000), application to the hierarchical analysis of spatial data (Banerjee et al. 2014) and to gene interactions (Friedman et al. 2004). Bayesian approaches are often combined with machine learning.

5. Use of machine learning: Machine learning techniques are also emerging as alternative methods for analyzing Big Data. Combining machine learning with statistical methods has yielded a new branch termed 'statistical learning' (Hastie et al. 2009). Supervised machine learning techniques are similar to traditional models (e.g. linear and logistic regression), which are based on predicting ecological phenomena (e.g., presence-absence of species, community-environment relationships, etc.) using quantifiable relationships generated from known data. Unsupervised machine learning techniques are designed to describe relationships in a manner similar to traditional clustering methods. Some supervised machine learning techniques, such as random forests (Breiman 2001) or boosted regression trees (Elith et al. 2008), have emerged as popular, viable alternatives to linear or additive mixed models. Regression tree-based methods can accommodate data that are autocorrelated, zero-inflated, non-monotonic, or otherwise difficult to fit into traditional modeling frameworks. Neural networks and support vector machines are also increasingly being used for statistical analyses (Folmer et al. 2016; O'Brien et al. 2016). Support vector machines have also been applied to automating the recognition of objects in huge sets of marine imagery (e.g. Beijbom et al. 2015). Deep machine learning methods, such as convolutional neural networks, are now also being used for this analysis (e.g. Luo et al. 2017). However, statistical and machine learning methods are often like a black box, and determining which method is most appropriate is often difficult.

These challenges highlight the importance of collaborations among ecologists, statisticians, and computer scientists. Advances in statistical computing have made computationally intensive machine learning models more accessible, allowing ecologists to spend less time on transforming and manipulating data and focus their attention on modeling complex ecological relationships.

4 Verification

Big Data is highly valued for its use in verifying and validating research, and improving confidence in ecological conclusions. It can be used to provide additional replicates, thereby improving the level of certainty in ecological conclusions, providing unprecedented opportunities in taxonomic surveys, co-occurrence studies, and habitat modeling. However, Big Data can be difficult to verify. Here we describe some methods for doing so, using examples from across aquatic ecology.

1. Amplicon sequencing for diversity assessment

Increasingly in microbiology, community composition and diversity are through the sequencing of highly conserved genes, such as 16S or 18S, a process that is also known as amplicon sequencing. However, methods used in the analysis of amplicon sequence data has the potential to lead to vastly different conclusions. These

differences largely stem from the clustering of sequences based on sequence similarity which is required for the identification of ‘consensus sequences’. Traditional methods have clustered sequences based on a set cut-off value which is provided by the researcher. Such approaches limit the ability of scientists to repeat the same results for a given sample, because the results are biased by the arbitrary starting point of the algorithm and the cut off set. Two new approaches address such a problem. First, the Swarm clustering algorithm addresses this problem by combining sequences using flexible cut offs to ensure that sequences are grouped appropriately, making results repeatable (Mahé et al. 2014). Second, there is a more recent push to move away from arbitrarily determined Operational Taxonomic Units and towards the use of Amplicon Sequence Variants (Callahan et al. 2017).

2. Remote sensing

Big Data acquired using remote sensing can provide large-scale coverage of the area surveyed in space and/or time. Verification is done by ground-truthing. For instance, seafloor maps generated by sonar can be ground-truthed by taking biogeophysical measurements (e.g., using cores) or by underwater imaging techniques such as sediment profile cameras or surficial benthic videos, which confirm sedimentary characteristics and provide information about the biological features of the area surveyed (Diaz et al. 2004). The level of certainty of such a seafloor map could be assessed by testing the level of agreement between classified areas of presumably homogeneous substrate types (e.g., muddy versus sandy areas) and co-occurring ground-truthed samples (e.g. cores).Techniques such as Cohen’s Kappa analysis (Cohen 1960) can be used to quantify the level of certainty in the classified map.

3. Statistical models using Big Data

It is essential to determine whether estimates of the variance explained by a given statistical model are realistic or “honest”. This concept goes beyond simply analyzing a dataset and estimating explained variance (r^2) to include cross-validation. Cross-validation uses an independent dataset to test whether a model produces results that are consistent if repeated (i.e., when applied to a different dataset or set of samples; Breiman et al. 1984; Flanagan and Cerrato 2015). Thus, cross-validation provides more realistic or “honest” estimates of the prediction error (uncertainty) of Big Data models.

4. Aquatic photography

The use of photography in aquatic ecology is rapidly increasing, and the task of detecting and classifying organisms in photographs is now being automated. Very large sets of images are difficult to verify manually, so a randomly chosen, representative, subset of data may be selected and validated to provide a measure of confidence in the whole dataset. A confusion matrix may be employed in this verification; it provides a tabular representation of the accuracy of automated versus manual classification, where mislabeling errors are easy to spot. This has been successfully applied in verifying the automated classification of plankton (Hu and Davis 2005) and fish (Shafait et al. 2016) in images.

5 Data sharing

Data sharing is a two-fold challenge for Big Data projects: (1) sharing large datasets once created and (2) sharing small datasets to be aggregated into Big Data. In aquatic ecology, it commonly involves the merging of environmental and biological datasets, often from different sources. Some suggest that ecologists already

collectively have sufficient data to tackle large-scale questions, if the data were aggregated (Hampton et al. 2013).

Data sharing culture is central to successful data sharing. Soranno et al. (2015b) notes that a culture for sharing publicly is lacking in the environmental sciences, and that a cultural shift towards data sharing must occur before it becomes common. They identify three current cultural barriers: a lack of rewards and incentives to share, greater risk in sharing for some groups of researchers (including early career researchers), and no ethical impetus for sharing in the culture. They also argue that this final barrier is the greatest to surmount, but that there is an ethical imperative to make data publicly accessible to make research inclusionary, and to facilitate good environmental policy development. Trust is central to the sharing culture, and it is often different between established colleagues than between unknown parties.

Barriers to data sharing still outweigh the benefits for many researchers (Soranno et al. 2015b), despite a multitude of benefits, including collaboration (Goring et al. 2014), increased citation, and discovery at larger scales. Reichman et al. (2011) found that just 1% of ecological data are accessible after the results have been published, while Nelson (2009) lamented the lack of data archived in repositories. Concerns about data sharing fall under three themes: ownership, control, and access. Borgman (2015a) describes the specific barriers as: “risks such as misuse, misinterpretation, liability, lack of experience, lack of tools and resources, lack of credit, loss of control, pollution of common pools of data, and the daunting challenges of sustainability”. Overcoming these barriers requires reflection and action at the individual researcher, institutional, and aquatic ecology community levels.

Another barrier to data sharing and reuse is the inability of researchers to see how their data may be useful to others, or valuable in the future (Borgman 2015a). Only a small proportion of aquatic ecologists use Big Data currently, but many, if not most, are collecting data that could be aggregated to form larger datasets. Thus, the consideration of potential future use is important for planning, collecting, and conserving data. However, this poses some challenges, including recognizing what constitutes data to other users. For example, what is considered to be metadata by one researcher may be primary data to another, so each researcher should curate their metadata as rigorously as their data. Some subgroups in aquatic ecology have used multidisciplinary discussion to posit the potential future uses of their data, and encourage data sharing even if such uses are currently unknown (e.g., Schoening et al. 2017). Nelson (2009) argued that data must be valued in equal measure to publications for sharing to occur; data publications may go some way to promoting data sharing without knowledge of its future use. We encourage researchers to consider the reuse of their data, for which Goodman et al. (2014) and Hart et al. (2016) present some practical advice, and to engage in such community discussions. Goodman et al. (2014) also provide a short guide to steps for data management that scientists can take to ensure their data continue to have value, such as making raw data available, publish methods/workflows, stating how credit should be given, and use of data repositories.

Challenges associated with the practicalities of data sharing are similar to those of data management, with the added challenge of multiple parties. Even in scientific domains where Big Data and the aggregation of data are common, data is commonly bartered (Wallis et al. 2013) and shared by email (Borgman 2015a). Such methods of data transfer are unsuitable for large datasets. If data is combined from several sources, all users need to trust it and its provenance before use; the increase in the velocity and volume of data generated with new technologies may exacerbate scepticism in data generated by others. Practically, the harmonization of data from different sources, and subsequent quality control requires computational skills and

substantial collaboration between parties, and is not an insignificant task (Soranno et al. 2015a).

We recommend that aquatic ecologists consider their data sharing culture, and the incentives and barriers associated with data sharing related to their own work – are they justified, and how could they be overcome? We suggest discussing possible future uses for data with colleagues both before and after the data are generated.

6 Developing knowledge infrastructure

Data management and sharing are facilitated by "knowledge infrastructure", a framework for interaction between researchers. Data repositories are the most-discussed type of knowledge infrastructure (Table 1). They are hubs of information, reducing the cost and effort of data maintenance for the individual researcher, and providing equitable and merit-based access to data (Borgman 2015a). They may be institutional or national, rather than domain-based (Goodman et al. 2014). Incentives to deposit data in repositories include requirements by funding agencies (e.g. European Commission 2017; National Science Foundation 2017), requirements for publication or encouragement by journals (e.g. Association for the Sciences of Limnology and Oceanography 2017; Ecological Society of America 2017; SpringerNature 2017), or precedent setting by societies (e.g. ASLO), as finding infrastructure funding for the common good is difficult (Nelson 2009). We recommend that aquatic ecologists consider depositing their data in a suitable repository, if possible.

Table 1. Examples of data repositories for aquatic ecology

Name	**Acronym/ Short name**	**Web page**
British Oceanographic Data Centre	BODC	bodc.ac.uk
Biological and Chemical Oceanography Data Management Office	BCO-DMO	bco-dmo.org
Dryad Digital Repository		datadryad.org
Environmental Data Initiative Data Portal	EDI Data Portal	portal.edirepository.org/nis/home.jsp
European Nucleotide Archive		ebi.ac.uk/ena
GenBank		ncbi.nlm.nih.gov/genbank
NOAA Environmental Research Division's Data Access Program	ERDDAP	coastwatch.pfeg.noaa.gov/erddap/
Ocean Biogeographic Information System		iobis.org
Publishing Network for Geoscientific and Environmental Data	PANGAEA	pangaea.de

Less discussed are the knowledge infrastructures related to data policy and culture. Data policies required for sharing or aggregation of data include decisions about what data is valuable, and how it should be shared, saved, curated, or maintained (Borgman 2015a). These data policies are initially made at the individual scientist level, but may be dictated by government, funding agencies, institutions, journal data

policies, or repositories. Such data policy is developed as a product of data culture, present at all levels. The development of a collective data policy in aquatic ecology may be challenging because of its interdisciplinary nature (Boyle 2013); researchers in different domains have different cultures around the collection, use, management, and sharing of data. It will require increased interdisciplinary collaboration within the community, for example exchange between biologists, engineers, and computer scientists. Knowledge infrastructures yet to be developed across the aquatic ecology community include an ontology, recognition of data and metadata diversity, structures for organising data, and defining the boundaries of data, which could be used to create policies, standards, and practises.

Knowledge infrastructures are difficult to design and maintain, particularly because of a lack of awareness about scientists' dependence on such infrastructure, and because maintenance involves invisible and undervalued work (Borgman 2015a). Furthermore, knowledge infrastructure development requires input and agreement by varied stakeholder groups, with failure likely if differences in theory, practice, and culture of data scholarship are not recognised. As such, creating infrastructure for research data sharing is a monumental task, and some suggest that it should be done at the research community level, rather than by non-profit organizations or universities (Nelson 2009). The development of a complex astronomical knowledge infrastructure involves long-term planning, regular coordination, and social, political and economic investment over decades and across continents (Borgman 2015a). Similar infrastructure has been developed for climate science, with the added incentive of the timeliness of data in this domain. The development of knowledge infrastructure to a similar degree in aquatic ecology is not a simple task, but individual researchers should consider the existing knowledge infrastructures and data culture within their research groups, institutions, and the broader aquatic ecology community, and be prepared to contribute to their development.

Conclusion

The six challenges and tools presented here highlight both practical and theoretical aspects of using Big Data in aquatic ecology, and include considerations for individual scientists and the wider community. We hope that these highlights provide some straightforward ideas to make interacting with Big Data more manageable, and to help readers reap the benefits offered by it. We also hope that they provide some topics for discussion between researchers and between research groups, not just on the culture necessary to foster the use of Big Data, but on collective initiatives and who should invest in them (Borgman 2015b). If we can overcome these challenges, aquatic ecology stands to benefit greatly from using Big Data, particularly in solving some of the largest environmental challenges facing our society.

Acknowledgements

This article is a result of discussions that began at the Eco-DAS XII: Ecological Dissertations in the Aquatic Sciences 2016 symposium. Thanks to the Center for Microbial Oceanography: Research and Education (C-MORE), the University of Hawai`i School of Ocean and Earth Science and Technology (SOEST) and the UH Department of Oceanography for hosting the event. Funding for EcoDAS XII was provided by the NSF biological oceanography program (Award OCE-1356192) and the Association for the Sciences of Limnology and Oceanography (ASLO). We thank C. Briseño-Avena, D.R. Miller, C. Filstrup and K.S. Cheruvelil for their comments that improved the manuscript.

References

Alexander, H., B. D. Jenkins, T. A. Rynearson, and S. T. Dyhrman. 2015. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. Proceedings of the National Academy of Sciences **112:** E2182-E2190.

Allcock, B. and others 2002. Data management and transfer in high-performance computational grid environments. Parallel Computing **28:** 749-771.

Association for the Sciences of Limnology and Oceanography. 2017. Limnology and Oceanography: Letters Author Guidelines, Data and Metadata Policy. http://aslopubs.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)2378-2242/about/author-guidelines.html.

Banerjee, S., B. P. Carlin, and A. E. Gelfand. 2014. Hierarchical modeling and analysis for spatial data. Crc Press.

Beijbom, O. and others 2015. Towards Automated Annotation of Benthic Survey Images: Variability of Human Experts and Operational Modes of Automation. PLoS ONE **10:** e0130312.

Benjamini, Y., and Y. Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society. Series B (Methodological) **57:** 289-300.

Bonferroni, C. E. 1935. Il calcolo delle assicurazioni su gruppi di teste. Tipografia del Senato.

Borgman, C. L. 2015a. Big Data, Little Data, No Data: Scholarship in the Networked World. The MIT Press.

Borgman, C. L. 2015b. If Data Sharing is the Answer, What is the Question? ERCIM News Special Theme: Scientific Data Sharing and Re-use **100:** 15-16.

Borgman, C. L. and others 2015. Knowledge infrastructures in science: data, diversity, and digital libraries. International Journal on Digital Libraries **16:** 207-227.

Boyle, J. 2013. Biology must develop its own big-data system. Nature **499:** 7.

Breiman, L. 2001. Random forests. Machine Learning **45:** 5-32.

Breiman, L., J. Friedman, R. Olshen, and C. Stone. 1984. Classification and regression trees. Wadsworth International Group.

Callahan, B. J., P. J. McMurdie, and S. P. Holmes. 2017. Exact sequence variants should replace operational taxonomic units in marker-gene data analysis. ISME J.

Cohen, J. 1960. A coefficient of agreement for nominal scales. Educational and Psychological Measurement **20:** 37-46.

Costa, B. M., T. A. Battista, and S. J. Pittman. 2009. Comparative evaluation of airborne LiDAR and ship-based multibeam SoNAR bathymetry and intensity for mapping coral reef ecosystems. Remote Sensing of Environment **113:** 1082-1100.

Diaz, R., M. Solan, and R. Valente. 2004. A review of approaches for classifying benthic habitats and evaluating habitat quality. Journal of Environmental Management **73:** 164-181.

Dray, S. and others 2012. Community ecology in the age of multivariate multiscale spatial analysis. Ecological Monographs **82:** 257-275.

Durden, J. M. and others 2016. Perspectives in visual imaging for marine biology and ecology: from acquisition to understanding, p. 1-72. *In* R. N. Hughes, D. J. Hughes, I.

P. Smith and A. C. Dale [eds.], Oceanography and Marine Biology: an Annual Review. CRC Press.

Ecological Society of America. 2017. Ecology Author Guidelines. http://esajournals.onlinelibrary.wiley.com/hub/journal/10.1002/(ISSN)1939-9170/resources/author-guidelines-ecy.html.

Elith, J., J. Leathwick, and T. Hastie. 2008. A working guide to boosted regression trees. Journal of Animal Ecology **77:** 802-813.

European Commission. 2017. Participant Portal H2020 Online Manual, Open access & Data management. http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm.

Fan, J., H. Fang, and H. Liu. 2014. Challenges of Big Data analysis. National Science Review **1:** 293-314.

Favali, P., L. Beranzoli, and A. de Santis. 2015. Seafloor Observatories: A new vision of the Earth from the Abyss. Springer-Verlag.

Fei, S., Q. Guo, and K. Potter. 2016. Macrosystems ecology: novel methods and new understanding of multi-scale patterns and processes. Landsc. Ecol. **31:** 1-6.

Flanagan, A., and R. Cerrato. 2015. An approach for quantifying the efficacy of ecological classification schemes as management tools. Cont Shelf Res **109:** 55-66.

Flanagan, A. M. 2016. Quantitative Benthic Community Models: The Relationship Between Explained Variance and Scale. Stony Brook University.

Folmer, E. and others 2016. Consensus forecasting of intertidal seagrass habitat in the Wadden Sea. Journal of Applied Ecology **53:** 1800-1813.

Friedman, N., M. Linial, I. Nachman, and D. Pe'er. 2004. Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology **7:** 601-620.

Gelman, A., and C. R. Shalizi. 2013. Philosophy and the practice of Bayesian statistics. British Journal of Mathematical and Statistical Psychology **66:** 8-38.

Goodman, A. and others 2014. Ten simple rules for the care and feeding of scientific data. PLoS Comput Biol **10:** e1003542.

Goring, S. J. and others 2014. Improving the culture of interdisciplinary collaboration in ecology by expanding measures of success. Frontiers in Ecology and the Environment **12:** 39-47.

Guisan, A., and N. E. Zimmermann. 2000. Predictive habitat distribution models in ecology. Ecol. Model. **135:** 147-186.

Hampton, S. E., L. R. Izmest'Eva, M. V. Moore, S. L. Katz, B. Dennis, and E. A. Silow. 2008. Sixty years of environmental change in the world's largest freshwater lake - Lake Baikal, Siberia. Global Change Biology **14:** 1947-1958.

Hampton, S. E. and others 2013. Big data and the future of ecology. Frontiers in Ecology and the Environment **11:** 156-162.

Hart, E. M. and others 2016. Ten Simple Rules for Digital Data Storage. PLoS Comput Biol **12:** e1005097.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Hobbie, J. E., S. R. Carpenter, N. B. Grimm, J. R. Gosz, and T. R. Seastedt. 2003. The US Long Term Ecological Research Program. BioScience **53:** 21-32.

Hooten, M. B., and N. T. Hobbs. 2015. A guide to Bayesian model selection for ecologists. Ecological Monographs **85:** 3-28.

Hu, Q., and C. Davis. 2005. Automatic plankton image recognition with co-occurrence matrices and Support Vector Machine. Marine Ecology Progress Series **295:** 21-31.

Jonsen, I. D., R. A. Myers, and J. M. Flemming. 2003. META-ANALYSIS OF ANIMAL MOVEMENT USING STATE-SPACE MODELS. Ecology **84:** 3055-3063.

Kitchin, R. 2014. Big Data, new epistemologies and paradigm shifts. Big Data & Society **1:** 2053951714528481.

Legendre, P. 1993. Spatial Autocorrelation: Trouble or New Paradigm? Ecology **74:** 1659-1673.

Legendre, P., and E. D. Gallagher. 2001. Ecologically meaningful transformations for ordination of species data. Oecologia **129:** 271-280.

Levy, O. and others 2014. Approaches to advance scientific understanding of macrosystems ecology. Frontiers in Ecology and the Environment **12:** 15-23.

Luo, J. and others 2017. Automated plankton image analysis using convolutional neural networks. Limnology and Oceanography: Methods **inpress**.

Mahé, F., T. Rognes, C. Quince, C. de Vargas, and M. Dunthorn. 2014. Swarm: robust and fast clustering method for amplicon-based studies. PeerJ **2:** e593.

Matabos, M. and others 2017. Expert, Crowd, Students or Algorithm: who holds the key to deep-sea imagery ‘big data’ processing? Methods in Ecology and Evolution **8:** 996-1004.

Mattmann, C. A. 2013. Computing: A vision for data science. Nature **493:** 473-473.

McGill, P. R., R. G. Henthorn, L. E. Bird, C. L. Huffard, D. V. Klimov, and K. L. Smith. 2016. Sedimentation event sensor: New ocean instrument for in situ imaging and fluorometry of sinking particulate matter. Limnology and Oceanography: Methods **14:** 853-863.

Morris, K. J. and others 2014. A new method for ecological surveying of the abyss using autonomous underwater vehicle photography. Limnology and Oceanography: Methods **12:** 795-809.

Müller-Wille, S., and I. Charmantier. 2012. Natural history and information overload: The case of Linnaeus. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences **43:** 4-15.

National Science Foundation. 2017. Chapter XI - Other Post Award Requirements and Considerations.

Nelson, B. 2009. Data sharing: Empty Archives. Nature **461:** 160-163.

O’Brien, C., M. Vogt, and N. Gruber. 2016. Global coccolithophore diversity: Drivers of future change. Progress in Oceanography **140:** 27-42.

Perez-Riverol, Y. and others 2016. Ten Simple Rules for Taking Advantage of Git and GitHub. PLoS Comput Biol **12:** e1004947.

Potts, J., and J. Elith. 2006. Comparing species abundance models. Ecol. Model. **199:** 153-163.

Reichman, O., M. Jones, and M. Schildhauer. 2011. Challenges and Opportunities of Open Data in Ecology. Science **331:** 703-705.

Schoening, T. and others 2012. Semi-Automated Image Analysis for the Assessment of Megafaunal Densities at the Arctic Deep-Sea Observatory HAUSGARTEN. Plos One **7:** e38179.

Schoening, T. and others 2017. Report on the Marine Imaging Workshop 2017. Research Ideas and Outcomes **3:** e13820.

Seiler, J., A. Williams, and N. Barrett. 2012. Assessing size, abundance and habitat preferences of the Ocean Perch Helicolenus percoides using a AUV-borne stereo camera system. Fisheries Research **129:** 64-72.

Shafait, F. and others 2016. Fish identification from videos captured in uncontrolled underwater environments. Ices J Mar Sci **73:** 2737-2746.

Smith, K. L., H. A. Ruhl, M. Kahru, C. L. Huffard, and A. D. Sherman. 2013. Deep ocean communities impacted by changing climate over 24 y in the abyssal northeast Pacific Ocean. Proceedings of the National Academy of Sciences **110**.

Sokal, R. R., and F. J. Rohlf. 1995. Biometry: the principles and practice of statistics in biological research. W.H. Freeman.

Soranno, P. A. and others 2015a. Building a multi-scaled geospatial temporal ecology database from disparate data sources: fostering open science and data reuse. GigaScience **4:** 28.

Soranno, P. A., K. S. Cheruvelil, K. C. Elliott, and G. M. Montgomery. 2015b. It's Good to Share: Why Environmental Scientists' Ethics Are Out of Date. BioScience **65:** 69-73.

Soranno, P. A., and D. S. Schimel. 2014. Macrosystems ecology: big data, big ecology. Frontiers in Ecology and the Environment **12:** 3-3.

SpringerNature. 2017. authors & referees > Policies > Availability of data, material and methods. Macmillan Publishers Limited.

Sullivan, G., and R. Feinn. 2012. Using effect size – or why the p value is not enough. Journal of Graduate Medical Education.

Swalwell, J. E., F. Ribalet, and E. V. Armbrust. 2011. SeaFlow: A novel underway flow-cytometer for continuous observations of phytoplankton in the ocean. Limnology and Oceanography: Methods **9:** 466-477.

Teytelman, L., A. Stoliartchouk, L. Kindler, and B. Hurwitz. 2016. Protocols.io: Virtual Communities for Protocol Development and Discussion. PLoS Biol **14:** e1002538.

Ver Hoef, J., and J. Jansen. 2007. Space—time zero-inflated count models of Harbor seals. Environmetrics **18:** 697-712.

Wagner, H. 2003. Spatial covariance in plant communities: Integrating ordination, geostatistics, and variance testing. Ecology **84:** 1045-1057.

---. 2004. Direct multi-scale ordination with canonical correspondence analysis. Ecology **85:** 342-351.

Wallis, J. C., E. Rolando, and C. L. Borgman. 2013. If we share data, will anyone use them? Data sharing and reuse in the long tail of science and technology. PLoS One **8:** e67332.

Wilson, G. and others 2014. Best practices for scientific computing. PLoS Biol **12:** e1001745.