The importance of open science for biological assessment

Marcus W. Beck ([marcusb@sccwrp.org](mailto:marcusb@sccwrp.org)), Raphael D. Mazor ([raphaelm@sccwrp.org](mailto:raphaelm@sccwrp.org)), Susanna T. Theroux ([susannat@sccwrp.org](mailto:susannat@sccwrp.org))

Version Date: Tue Oct 9 13:10:06 2018 -0700

# Abstract

Open science principles that seek to democratize science can effectively bridge the gap between researchers and environmental managers, yet widespread adoption has yet to gain traction for the development and appplication of bioassessment methods. At the core of this philosophy is the concept that research should be reproducible and transparent, in addition to having long-term provenance through effective modes of data preservation and sharing. We discuss core open science concepts that have been advocated more generally in the ecological sciences and will emphasize how adoption can benefit bioassessment for both prescriptive condition assessments and proactive applications that inform planning activities. Examples from the state of California will be used to demonstrate effective adoption of open science principles through data stewardship, reproducible research, and engagement of stakeholders with multimedia applications. Technical, sociocultural, and institutional challenges for adopting open science will also be discussed, including practical approaches for overcoming these hurdles in bioassessment applications.

# Introduction

Bioassessment is an essential element of aquatic monitoring programs that establishes a foundation of decisions for managing the ecological integrity of environmental resources. Legal mandates to assess biological condition have set a precedent for developing bioassessment methods in the United States (Clean Water Act, CWA), Canada (Canada Waters Act), and Europe (Water Framework Directive). Decades of research to meet these mandates have supported the development of methods for multiple assemblages with regional applications in streams, rivers, lakes, and marine environments (Karr et al. [1986](#ref-Karr86); Kerans and Karr [1994](#ref-Kerans94); Fore and Grafe [2002](#ref-Fore02); Beck and Hatch [2009](#ref-Beck09); Borja, Ranasinghe, and Weisberg [2009](#ref-Borja09)). This body of applied tools represents substantial progress in understanding how biological organisms can be used as accurate and interpretable sentinels of environmental condition. Monitoring programs in the United States and internationally have collected millions of records of biological data spanning decades and hundreds of assessment methods have been developed from these data. The explicit link to environmental management distinguishes bioassessment from basic ecological research. Although bioassessment can and has been used to inform basic research, the intended use of these tools is to inform the protection and restoration of ecological integrity.

Bioassessment products will have limited use if they do not meet the needs of management and regulatory communities (Bain et al. [2000](#ref-Bain00); Stein et al. [2009](#ref-Stein09); Kuehne et al. [2017](#ref-Kuehne17)). In the United States, the CWA gives power to states, tribes, and territories for method development, which in turn requires federal approval to be implemented in a regulatory framework(e.g., Total Maximum Daily Load reporting, stormwater permitting). An imbalance exists between the developed methods and those that are approved for regulatory use. A recent review of assessment methods for ecological integrity in the US showed that few were explicitly connected to freshwater policy (Kuehne et al. [2017](#ref-Kuehne17)). Of those methods that are actively used, a more problematic issue is the manner of application within standard regulatory frameworks. Biological indices are typically used to develop post-hoc diagnoses to trigger remediative or restoration actions, or serve as early warning indicators of environmental change (Niemi and McDonald [2004](#ref-Niemi04)). A critical concern is that these tools, although technically sound, are implicitly being used to document the long-term demise of environmental health. A much broader use for bioassessment to guide planning activities, such as identifying conservation priorities (Linke, Turak, and Nel [2011](#ref-Linke11); Howard et al. [2018](#ref-Howard18)), could extend the reach of bioassessment products beyond regulatory applications.

An effective bioassessment product must jointly address the technical challenges of developing a diagnostic index and the implementation challenges of using an index to address the needs of environmental managers and regulators (Jackson and Davis [1994](#ref-Jackson94); Dale and Beyeler [2001](#ref-Dale01)). For decades, research in the bioassessment community has focused on addressing the former. Substantial technical advances have been made in predicting biological responses to environmental change, how these responses can be distinguished from natural environmental variation, and determining the impacts of these changes. Standardized protocols have also been developed (McDonald et al. [2004](#ref-McDonald04); Stoddard et al. [2008](#ref-Stoddard08)). Many bioassessment indices are characterized as either multimetric, such as an index of biotic integrity (Karr et al. [1986](#ref-Karr86)), or multivariate where condition is based on predictive methods that evaluate similarity of taxonomic composition to reference expectations. The reference-condition approach establishes the foundation for many bioassessment methods whereby a set of reference sites are identified and used to evaluate levels of biological deviation to define potential impacts (Reynoldson et al. [1997](#ref-Reynoldson97); Stoddard et al. [2006](#ref-Stoddard06)). Most indices can be broadly categorized in the context of these well-established technical methods.

Implementation challenges have severely limited the use of bioassessment products in management and regulatory applications relative to the availability of developed methods. Characterizing how an index could be used in practice to inform decisions and prioritize management actions is often opaque relative to why an index may have been originally developed. Hundreds of assessment methods have been developed for specific regional applcations (Birk et al. [2012](#ref-Birk12)) and concerns about redundancy, comparability, duplicated effort, and lack of coordinated monitoring have recently been discussed within the research community (Cao and Hawkins [2011](#ref-Cao11); Poikane et al. [2014](#ref-Poikane14); Kelly et al. [2016](#ref-Kelly16); Nichols et al. [2016](#ref-Nichols16)). Moreover, existing methods may not be discoverable beyond immediate research applications (Hering et al. [2010](#ref-Hering10); Nichols et al. [2016](#ref-Nichols16)) or may be incorrectly applied based on differences between goals for developing an index and the needs of management programs (Dale and Beyeler [2001](#ref-Dale01); Stein et al. [2009](#ref-Stein09)). The abundance of available methods can be a point of frustration for managers given a lack of guidance for choosing an appropriate method among alternatives, particularly as to how a method may relate to specific management, monitoring, or policy objectives (Dale and Beyeler [2001](#ref-Dale01); Stein et al. [2009](#ref-Stein09)).

Environmental managers require additional tools that transform bioassessment data and methods into actionable information. A new mode of operation is needed where method development is open and transparent, developed products are discoverable and reproducible, and most importantly, implementation in the management community is intuitive and purposeful. Open science principles that can democratize all aspects of the scientific method can meet these needs, yet bioassessment research and its application to better serve the environment has not fully embraced these principles. Others have advocated more broadly for inclusion of open science principles in the ecological sciences (Hampton et al. [2015](#ref-Hampton15), [2016](#ref-Hampton16); Lowndes et al. [2017](#ref-Lowndes17)) and a growing wave of momentum has influenced how scientists conceptualize research in other disciplines (e.g., archeaology, Marwick et al. [2016](#ref-Marwick16); behavioral ecology, Ihle et al. [2017](#ref-Ihle17); vegetation sciences, Collins [2016](#ref-Collins16)). Adopting an open science paradigm in biaossessment is particularly relevant compared to other fields given the explicit need to develop tools that are open and accessible to the management community. Legal and ethical precedents in bioassessment may also necessitate the open sharing of data given that environmental monitoring programs are often publicly funded.

This review will demonstrate tools and approaches for open science, which will empower the research and management community to embrace a new mode of thinking for bioassessment applications. These approaches are expected to benefit the research community by augmenting existing workflows for developing assessment tools, but more importantly, improve the ability of these methods to address environmental issues by bridging the gap between the scientific, management, and regulatory communities. An overview of the general principles of open science is provided, followed by a discussion of specific benefits and how these principles can be applied to bioassessment. We use examples from the state of California to demonstrate how existing tools can be tailored to address legislative mandates for free and open sharing of data, especially by directly engaging stakeholders that require practical approaches for using bioassessment tools in planning activities. We conclude with a discussion of technical, sociocultural, and instutional hurdles that have, thus far, prevented widespread adoption of open science and provide recommendations for the bioassessment community to address these challenges.

# Principles of open science

Conventional modes of creating scientific products and more contemporary approaches that align with open science principles share the same goals. Both are motivated by guiding principles of the scientific method that make the process of discovery transparent and repeatable. Where the conventional and open science approaches diverge is the extent to which technological advances are leveraged as instrumental tools that facilitate the entire research process. Distinction between the two approaches can be conceptualized as the “paper as the only and final product” for the conventional approach, whereas the open science approach is inherently linked to advances in communication and analysis that have been facilitated by the Internet and computer sciences. As a result, the open science approach can enhance all aspects of the scientific process from initial conception of a research idea to the delivery and longevity of a research product (Figure 1). The process is iterative where products are improved by the individual and/or others, facilitated by open science tools that enhance access and reproducibility of data.

The paradigm of the scientific paper as a final research product can inhibit forward progress for several reasons, particularly so in the applied sciences. The research paper is conventially viewed as a communication tool for scientists to report and share results among peers. Researchers access periodicals to stay informed of scientific advances and use the information to replicate methods for follow-up analysis. Although the primary literature continues to provide these fundamental services, this workflow is problematic when scientific products are needed to serve interests outside of the research community. For example, the paper as an endpoint for environmental managers fails to deliver tools that are easily accessible from the practitioners perspective, both in application and interpretation. A research paper rarely affects environmental change because it does not provide a mechanism to transfer actionable information to those that require scientific guidance for decision-making. Numerous studies have documented implementation failures as a result of siloing among research communities where the flow of information does not extend beyond institutional walls. The loss of information over time that intimately describes a research product is another well-known flaw associated with the paradigm of research paper as final product (Michener et al. [1997](#ref-Michener97)).

The open science approach makes the researcher a steward of their data. This allows the data to be treated as a living product with a history, rather than proprietary and serving only the internal needs of an immediate research goal. Data can be generically described as any component of the research process that is used to address a research goal and could include literal tabular data, a laboratory notebook, a research report, data visualizations or maps, analysis code or software, or even presentation materials. Open science principles can be applied to any and all of these data with the end goal of facilitating communication for researchers and those for which the research was developed (Bond-Lamberty, Smith, and Bailey [2016](#ref-BondLamberty16)). These tools also benefit the individual researcher by providing information for the “future-self” to recreate a past working environment. In all cases, the data are openly accessible and documented for reproducibility and discovery using technologies that facilitate communication and sharing.

Open data can benefit research by contributing to an increase in novel products created through collaboration. Collaborative publications have increased in the environmental sciences as researchers leverage open data to create synthesis products as the sum of individual datasets. Quantitative meta-analyses and systematic reviews are increasingly used to extract information from the primary literature (Lortie [2014](#ref-Lortie14)). In addition, open data products can increase efficiency of the individual researcher and a collective research team by encouraging collaborators to adopt an open science workflow. Many tools developed within the software and computer science community are now easily accessible to environmental scientists that can be used to create open data. Version control software (e.g, Git, GitHub), open source programming languagues (e.g, R), and integrated development environments (IDEs, e.g., RStudio) can all be leveraged by applied ecologists to dynamically create and share open data products that build instutional memory. These tools promote deliberate and shared workflows among researchers that can lead to better science in less time (Lowndes et al. [2017](#ref-Lowndes17)).

Open and unfettered access to data can also benefit management and regulatory communities. Although many monitoring datasets are publicly available, the data quality, level of documentation, and ease of use varies widely. This can create research and implementation challenges as datasets may be difficult to locate, contexts for data may be misunderstood, and data from different sources require synthesis. Many open science tools can improve the accessibility of data from monitoring programs by establishing workflows for data synthesis and discovery, often through the adoption of a common metadata structure (e.g., Ecological Metadata Language Standard) and integration of data within federated data networks (e.g., DataONE). Open data maintained by management or regulatory communities benefits the research community, which in turn benefits the data maintainers that require scientific products to inform decisions. Open data can also improve public trust in scientific findings by exposing the underlying information used to develop a research product (Grand et al. [2012](#ref-Grand12)). Increased trust could facilitate eventual adoption of proposed rules or regulations that are based on research products created from open data.

# Applying open science principles to bioassessment

Here we provide a detailed description of science tools that the bioassessment community could leverage to adopt a a philosophy of creating reproducible, transparent, and discoverable research products for environmental managers. To emphasize the value that each of these tools can have for specific steps of the scientific process, we first describe a “conventional” workflow, which is then constrasted with a workflow that adopts open science tools. In both cases, the technical and implementatation phases of a bioassessment product are acknowledged as distinct steps of analysis that describe the entire process from idea conception to adoption in management or regulatory applications. We use this paradigm to jointly demonstrate how open science tools can be applied beyond the research phase, but also reinforce the concept that a bioassessment product is only as relevant as its applied context (i.e., an index will not have value if its final home is the primary literature).

## Closed science

First we consider a hypothetical workflow for how many bioassessment products are potentially created (Figure 2a). Research goals are identified through legislative requirements to assess biological integrity of navigable surface waters. In the United States, this need originates at the state-level where regulatory institutions require robust bioassessment methods. Historically, a regulatory agency may not have ownership of the data required for index development, as for states that separately manage pollutants (e.g., stormwater or discharge permits) and natural resources (e.g., fisheries). The required data may be decentralized across public and private institutions. The task of identifying, gathering, and synthesizing these data may be contracted by the state regulatory agency to a third party, such as a private consultant or an academic institution. This separation between consumers and creators of bioassessment products can be an initial cause of implementation challenges as research agendas may diverge with communication barriers between institutions.

A typical workflow for developing an index is not entirely dissimilar from a conventional scientific process. The primary investigator begins by identifying the research goal, developing methods to achieve that goal, and then identifying the data needs and analyses based on the existing sciencd. Development of a bioassessment index is not explicitly hypothesis-driven, in that an index is not meant to support or refute an academic question, but rather it is implicitly guided by the needs of an ecological condition assessment. The index must predict biological responses to environmental change relative to natural variation and the relative impacts of these changes should be quantifiable. As noted above, standard procedures that address these technical challenges have generally been accepted by the research community as robust and index development could be considered more procedural than research-oriented. This can contribute to a relaxation of scientific principles that reinforce reproducibility and attention to detail if the development process is viewed more as a recipe than a formal research endeavor.

Inadequate documentation of data sources, what data were used, and how data were synthesized can also contribute to implementation challenges in conventional workflows. Ideally, a dataset from a regional monitoring program would be available that covers the range of natural variability and relevant stressor gradients. If so, the origin can easily be identified after the product is developed and others will more easily be able to recreate or refine as needed. However, a pre-existing monitoring program may not have goals that align with the needs of a bioassessment product. The data may also have limitations that affect performance or even prevent their use entirely for assessment. For example, a fisheries monitoring program may only sample select locations and may have inadequate coverage of aquatic biodiversity. In these scenarios, supplementary data must be gathered from other sources, all of which potentially have different goals for originally collecting the data. The end product is a synthesized dataset that is wrangled to meet the requirements of a bioassessment product, but its origin and workflow to achieve the necessary format may be unknown except to the individual that compiled the data. This creates significant challenges for future products or policies that depend on the product and may even erode institutional trust if the provenance of the dataset is unknown.

Finally, most bioassessment workflows usually lack tools for calculation and interpretation of results. Although the availability and need for documentation of the raw data is important, insufficient resources to facilitate use by managers and policy-makers, both in calculation and interpretation, represents the primary limitation of bioassessment effectiveness with conventional workflows. The worst case scenario is a spreadsheet-oriented approach to converting raw data into a synthesized index. Not only is this a major challenge for the researcher that develops the index, but it prevents others from applying the tool to reproduce the results and apply it with novel data. There are many reasons for these challenges (e.g., lack of documentation, unknown data requirements, no interpretation guidance), but the fundamental issue is inaccessibility of the product. A management or regulatory agency will have no ability to interpret these data and the final product will not pass independent review if others cannot recreate the results.

## Using open science tools to enhance bioassessment

The above example is an extreme hypothetical case where bioassessment products fail to affect any positive change in environmental management due to a complete deficiency in openness and failed implementation as a result. Many bioassessment methods can be described more positively, yet the research community could benefit from adopting a more open approach to creating and delivering assessment tools. This approach is particularly relevant from the perspective of implementation and science translation, as all modes of communication between research, managers, and regulators could be enhanced with open science. The following section describes key components of the open science process that can faciliate the development and implementation of bioassessment products for affecting positive environmental change. We focus specifically on open science applications for data provenance, method development, and method delivery.

The overall process is shown in Figure 2b as an expansion of general concepts in Figure 1, with a specific science translation phase for implementation. The critical difference of this open approach with the closed scenario in Figure 2a is the iterative flow of ideas and products between the management community and stakeholders, the researcher developing the bioassessment product, and the broader community that provides data and guidance documents in the primary and secondary literature. This iterative flow of information is facilitated by 1) openly sharing planning documents, 2) using established metadata standards to document synthesized data products, 3) hosting data products on open repositories, 4) creating reproducible summary documents that integrate the data and research products, and 5) incorporating the developed product into interactive applications that deliver the results to the stakeholders. The technical phase of defining research goals, collecting and synthesizing data, and developing the bioassessment product remains the sole responsibility of the researcher. However, the process is distinguished by the flow of information to and from the research phase that can benefit the specific project and the science of bioassessment as a whole.

### Developing bioassessment goals

In an open science paradigm, the goals identified by the researcher for developing a bioassessment product should occur through close interaction with the management or regulatory institution that requires the product. Establishing goals should occur through a two-way exchange of information where the regulatory institution communicates the assessment needs of a bioassessment product that reflects both legislative reporting requirements and stakeholder concerns. The researcher developing the bioassessment product should recognize these needs while also considering the potential balance between the goals and limitations of the data or state of the science to meet these goals. This two-way exchange of information can be accomplished through direct communication and sharing planning documents to ensure all decisions are transparent. In person meetings are ideal, but planning documents are dynamic and will require remote sharing and revision as ideas progress. Online tools such as Google documents, Slack discussion channels, and open lab notebooks can be instrumental for collaboration. More informal approaches, such as blogging and sharing ideas on social media, can expose new concepts to the broader community for guidance. Overall, the iterative exchange of information for identifying goals will ensure that the needs of the management and stakeholder communities will be consistent with the services provided by the research product.

### Curating bioassessment data

After the goals are established, the researcher identifies requirements and sources of data that need to be synthesized to meet the research needs. Under a closed scenario, data flows one way from the source to the researcher and is used only as a means to create the final research product. In the open scenario, the data itself is a product to achieve the research goals and also becomes available to the research and management community as a fully documented source of information that has value beyond the specific project. The openness of the synthesized data product is one of the primary means of facilitating the implementation phase of a bioassessment product. The synthesized data product can be used by the individual researcher to create interactive applications for stakeholders to share and explore the data and is also fully integrated into summary reports using software for generating dynamic documents (e.g, using knitr, Xie [2015](#ref-Xie15); and RMarkdown, Allaire et al. [2018](#ref-Allaire18)). The data product also becomes available on an open data repository that is discoverable by other researchers and can contribute to alternative scientific advances beyond the immediate goals.

A bioassessment researcher operating under an open science paradigm has the responsibility of curating the data from its initial creation to its final home in an open repository. This responsibility is particularly relevant given that a synthesized dataset can originate from multiple sources and the created output can be indistinguishable from the original data. For example, a multimetric index may require taxonomic data collected at multiple sites by different institutions, whereas the output data may include summary scores, individual metrics, and any additional supporting information to assess the quality of the output. This requires use of a standardized metadata language (e.g., the Ecological Metadata Language or EML) to document the who, what, and why of a particular dataset. Adoption of a metadata standard also ensures that a machine-readable file is produced to allow integration into a data repository. This will allow a synthesized data product to be discoverable beyond the specific research application and will include the metadata to understand the context of the data. Finally, a dataset can be assigned a unique digial object identifier (e.g., through Zenodo) that provides a permanent address and is also citable to allow researchers to track usage of a bioassessment data product.

### Using R for bioassessment translation

The most important process of the open science workflow is the translation of bioassessment products to the management and regulatory community. This process should be fundamentally linked to open source analysis and development tools that can be used to deliver the products using a reproducible and accessible platform. In particular, the popularity of the R statistical programming language (RDCT (R Development Core Team) [2018](#ref-RDCT18)) has increased in the last ten years and is the most commonly used analysis platform in the environmental sciences. This software provides 13134 user contributed packages (as of 2018-10-11) and is also a programming language that can be used to create specific analysis workflows. The availability of existing packages and the ability to create new packages is a strength of R that is under-utilized by the bioassessment community. An assessment index packaged in R can automate the tedious process of converting raw taxonomic data to summary scores. An R package is also modular, meaning it includes all necessary analyses, data, and documentation to allow use by others. This modularity is critical for reaching management and stakeholder communities by providing a tool that redirects the focus on the output and not the technical details that are less important for decision-making.

Several existing R packages have value for the bioassessment community (Table 1). For managing the day to day tasks of working with multiple datasets, the tidyverse suite of packages provides the necessary tools to import, wrangle, explore, and plot almost any data type (Wickham [2017](#ref-Wickham17b)). These packages are developed around the concept of “tidy” data that provide a common and natural framework for working with data (Wickham [2014](#ref-Wickham14c)). These principles have importance for bioassessment where the synthesized datasets used to create a product should be logically structured to facilitate use by others. Although the time and effort required to produce a tidy dataset can seem excessive relative to native formats (e.g., raw sample data), downstream analyses will be greatly facilitated. The tidyverse also includes the popular ggplot2 package that is based on a syntactical grammar of graphics for plotting (Wilkinson [2005](#ref-Wilkinson05); Wickham [2009](#ref-Wickham09)). This package provides a set of independent plotting geometries and aesthetics that can be built piecewise and is a departure from other graphics packages that represent a collection of special cases that limit the freedom of the analyst. In bioassessment, ggplot2 can be used both in an exploratory role during the development phase and also to create publication quality graphics. More importantly, this package provides the building blocks to create a data visualization to convey important components of a bioassessment product to managers and stakeholders.

Bioassessment data are inherently spatial and recent package development has greatly improved the ability to analyse and map geospatial data in R. These tools can readily communicate the spatial context of bioassessment products to managers and stakeholders by mapping index or condition scores to stream flow networks, watersheds, and ecoregions, both for high-level planning and site-specific evaluations. The simple features package (Pebesma [2018](#ref-Pebesma18)) was first released in 2016 and has quickly become the most popular approach for working with spatial information in R. The simple features package uses principles of data storage that parallel those from the tidyverse by representing spatial objects in a tidy and tabular format. This facilitates analysis by presenting complex spatial structures in an easily readable format that can integrated in workflows with existing packages, including other mapping packages (e.g., leaflet, Cheng, Karambelkar, and Xie [2018](#ref-Cheng18); or mapview, Appelhans et al. [2018](#ref-Appelhans18)). This allows the researcher to use a workflow that is focused in a single environment, rather than using separate software for statistical and geospatial analysis. This reduces the likelihood of losing information or introducing analysis errors as the workflow is more transparent in a single environment.

Other existing R packages can be used to develop statistical models of bioassessment data that are a necessary component of many analyses. Random forest models have been used to develop predictive bioassessment indices that compare observed taxa to modelled expectations (i.e., O/E indices). The randomForest package (Liaw and Wiener [2002](#ref-Liaw02)) uses an ensemble learning approach that is robust to complex, non-linear relationships and interactions between variables. These models are particularly useful with large, regional datasets that describe natural and anthropogenic gradients in condition. Many other modelling packages are available in R that can support index development, such as exploratory analyses to evaluate biological response or identifying significant associations of organisms with stressor gradients. The nlme package can be used to create non-linear mixed effect models that are more flexible than standard regression approaches (Pinheiro et al. [2018](#ref-Pinheiro18)). The nlme package can develop models for nested sampling designs, such as repeat visits to sample sites or otherwise confounding variables that contribute information but are not unique observations. The mgcv package provides similar functionality as nlme, but uses an additive modelling approach where invididual effects can be evaluated as the sum of smoothed terms (Wood [2017](#ref-Wood17)).

Currently, only a few existing R packages have been developed speicifically for bioassessment. For example, the TITAN2 package can be used to developed quantitative evidence of taxon-specific changes in abundance and occurrence across environmental gradients (Baker, King, and Kahle [2015](#ref-Baker15)). Results from this package can support exploratory analysis for developing bioassessment products, such as identifying indicator species. The results can be also be used post hoc to evaluate potential response of a biological index with changing environmental conditons, such as proposed management actions to improve aquatic habitat. Alternatively, the indicspecies package provides similar functionality but is based only on species occurrence or abundance matrices across sites (De Caceres and Legendre [2009](#ref-DeCaceres09)). This package can be used to identify species that occur at particular sites if continuous environmental data are unavailable, such as those that are representative of reference conditions. Finally, the vegan package has been a staple among community ecologists for multivariate analyses in R, such as clustering and ordination (Oksanen et al. [2018](#ref-Oksanen18)). This package has value for bioassessment as an exploratory tool with a variety of applications, such as analyses that identify comparable assessment sites that minimize natural variation (Beck, Vondracek, and Hatch [2013](#ref-Beck13)).

Although the R network includes over 10000 user contributed packages, the most underutilized aspect of this software for bioassessment is the creation of new packages to support implementation. Several tools have been developed and published in the last five years that simplify the process of creating new packages in R (Wickham [2015](#ref-Wickham15); Wickham, Hester, and Chang [2018](#ref-Wickham18)). The increasing popularity in the environmental sciences of online repositories for hosting code and software, such as GitHub, also provides a venue for sharing these packages. The advantages of creating and sharing R packages that are specific to bioassessment applications are important for several reasons. First, an R package creates a compartmentalized set of instructions developed during the technical research phase that can be executed by anyone with access to a computer. This allows the developer to include important technical elements required for the execution of a bioassessment product within the package, while allowing the end user to focus on how the output can be used to inform decision-making. R packages also require explicit documentation of the functions and data requirements. As such, package users will not only have access to underlying code but also understand the why and what for different package functions. Detailed vignettes can be included in plain language to describe how to use the package.

Finally, R can be used to create interactive applications that deliver bioassessment products to stakeholders and managers in entirely novel contexts. In particular, the shiny package was first released for R in 2012 and provides programming tools built around concepts of reactivity, where inputs and outputs can be modified in real time (Chang et al. [2018](#ref-Chang18)). A shiny application is an interactive user interface that is developed with R code, but is a standalone product that can be used without any programming experience. These applications are deployed online and can extend the reach of bioassessment tools to those that require the information for decision-making but otherwise do not have the time or resources to learn R. Shiny applications can also be easily linked to other R packages. For example, a shiny website could be created for users to upload raw data and bioassessment scores could be estimated and reported using an R package developed externally. This can extend the accessibility of a bioassessment product while maintaining the technical integrity of the original tool. Moreover, shiny applications are completely customizable and can be tailored to the specific needs of any users. This distinction separates shiny from other web-based analysis platforms.

# Case studies of open science in practice

# Limitations and solutions

Although the case for open science in bioassessment is appealing, several challenges have prevented the widespread adoption of these principles in practice. Many of these challenges are not specific to bioassessment and apply more generally to the environmental sciences. First, the technical skillsets required to effectively use open science tools can be prohibitive for some researchers. Conventional and closed workflows used by many scientists are adopted and entrenched because of ease of use, yet the simplicity of such tools limits their power. In contrast, many open science analysis platforms, such as R, were developed with customization as a key concept defining the functionality. This creates a tradeoff between ease of use and flexibility where the former is sacrificed by the latter. Open science tools can improve analysis, documentation, and implementation through greater flexibility but they expose researchers to entirely new concepts in which they were never trained. Many scientists are unwilling to learn new skills given existing demands on their time, and even less so for skills with exceptional learning curves. The benefits of these approaches may also not be apparent, such as the value for the individual researcher of sharing their data. Not only are the required skillsets demanding, but the open science toolbox continues to expand as new methods are developed and old methods become obselete. This requires a researchers to stay abreast of new technologies as they are developed and weigh the tradeoffs of adopting different workflows for different research tasks.

Advocates for open science are well aware of the technical challenges faced by individuals that have never been exposed to the core concepts and several approaches have been used to foster acceptance. Most importantly, education and training remain key components for developing skillsets among researchers where the focus is both on learning new skills and realizing their value for improving science as a whole. A goal of many training curriculums is to instill confidence in new users by developing comfort with new workflows, such as replacing a point-and-click style of analysis with one focused on using a command line through a computer terminal. Other approaches to demonstrate the value of new techniques use a side by side approach of closed vs open workflows to show the increased efficiency and power of the latter. Adoption becomes much more feasible once the value of investing in learning a new skill is realized. Advocates of open science also realize the limitations of teaching in that not all audiences can be reached and not all materials are retained or even used after training. A technique of letting the trainee become the trainer can be used whereby those that successfully learn new skills can teach others at their home institutions. Those that also adopt new workflows through training can also direct their research products to facilitate collaboration with non-adopters rather than the latter, for example, analyzing their data in potentially subopotimal ways.

Sociocultural or institutional challenges may also prevent the adoption of open science methods. The most common argument againt open science is the protection of data products that

# Figures



Figure 1 A simplified workflow of the open science paradigm (adapted from Hampton et al. ([2015](#ref-Hampton15))). All aspects of the research process, from the conception of an idea to publishing a product, can be enhanced using open science tools. The workflow is iterative where products are continually improved through collaborations facilitated through discovery and reproducibility of open data.

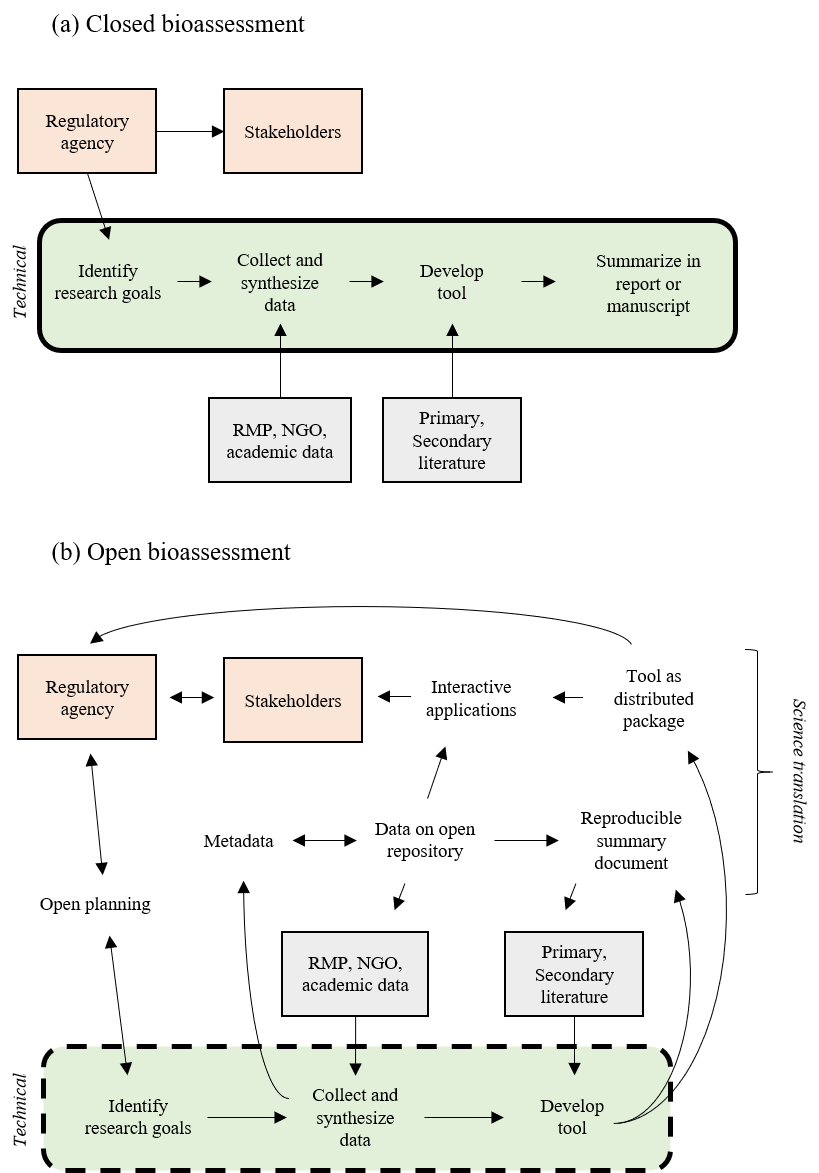


Figure 2 Examples of (a) closed and (b) open approaches for bioassessment applications. The oval box in each subfigure represents the technical steps of the individual researcher for developing the product, the regulatory and stakeholder boxes indicate those that require or motivate the creation of bioassessment products, the gray boxes indicate sources of external information (data and guidance documents) as input into the technical process, and the open text indicates open components of the planning or implementation phase of a bioassessment product. Figures were adapted from Hampton et al. ([2015](#ref-Hampton15)). NGO: non-government organization, RMP: regional monitoring program.

# Tables

Table 1 R packages that can be used in the development, application, and translation of bioassment products.

|  |  |  |
| --- | --- | --- |
| Task | Package | Description |
| General | tidyverse (Wickham 2017) | A suite of packages to import, wrangle, explore, and plot data. Includes the popular ggplot2 and dplyr packages. |
| Mapping, geospatial | sf (Pebesma 2018) | A simple features architecture for working with spatial data, including common geospatial analysis functions |
|  | leaflet (Cheng et al. 2018) | Integration of R with the popular JavaScript leaflet library for interactive maps |
|  | mapview (Appelhans et al. 2018) | Creates interactive maps to quickly examine and visually investigate spatial data, built of leaflet and integrated with sf |
| Statistical modelling | randomForest (Liaw and Wiener 2002) | Create classification and regression trees for predictive modelling |
|  | nlme (Pinheiro et al. 2018) | Non-linear, mixed effects modelling |
|  | mgcv (Wood 2017) | Generalized additive modelling |
| Community analysis | TITAN2 (Baker et al. 2015) | Ecological community threshold analysis using indicator species scores |
|  | indicspecies (De Caceres and Legendre 2009) | Indicator species analysis |
|  | vegan (Oksanen et al. 2018) | Multivariate analysis for community ecology |
| Science translation | shiny (Chang et al. 2018) | Reactive programming tools to create interactive and customizable web applications |
|  | rmarkdown (Allaire et al. 2018) | Tools for working with markdown markup languages in .Rmd files |
|  | knitr (Xie 2015) | Automated tools for markdown files that process integrated R code chunk |

# References

Allaire, JJ, Yihui Xie, Jonathan McPherson, Javier Luraschi, Kevin Ushey, Aron Atkins, Hadley Wickham, Joe Cheng, and Winston Chang. 2018. *Rmarkdown: Dynamic Documents for R*. <https://CRAN.R-project.org/package=rmarkdown>.

Appelhans, Tim, Florian Detsch, Christoph Reudenbach, and Stefan Woellauer. 2018. *Mapview: Interactive Viewing of Spatial Data in R*. <https://github.com/r-spatial/mapview>.

Bain, M. B., A. L. Harig, D. P. Loucks, R. R. Goforth, and K. E. Mills. 2000. “Aquatic Ecosystem Protection and Restoration: Advances in Methods for Assessment and Evaluation.” *Environmental Science and Policy* 3 (1):89–98.

Baker, Matthew E., Ryan S. King, and David Kahle. 2015. *TITAN2: Threshold Indicator Taxa Analysis*. <https://CRAN.R-project.org/package=TITAN2>.

Beck, M. W., and L. K. Hatch. 2009. “A Review of Research on the Development of Lake Indices of Biotic Integrity.” *Environmental Reviews* 17:21–44. <https://doi.org/10.1139/A09-001>.

Beck, M. W., B. Vondracek, and L. K. Hatch. 2013. “Environmental Clustering of Lakes to Evaluate Performance of a Macrophyte Index of Biotic Integrity.” *Aquatic Botany* 108:16–25. <https://doi.org/10.1016/j.aquabot.2013.02.003>.

Birk, S., W. Bonne, A. Borja, S. Brucet, A. Courrat, S. Poikane, A. Solimini, W. van de Bund, N. Zampoukas, and D. Hering. 2012. “Three Hundred Ways to Assess Europe’s Surface Waters: An Almost Complete Overview of Biological Methods to Implement the Water Framework Directive.” *Ecological Indicators* 18 (1):31–41. <https://doi.org/10.1016/j.ecolind.2011.10.009>.

Bond-Lamberty, B., A. P. Smith, and V. Bailey. 2016. “Running an Open Experiment: Transparency and Reproducibility in Soil and Ecosystem Science.” *Environmental Research Letters* 11 (8):1–7. <https://doi.org/10.1088/1748-9326/11/8/084004>.

Borja, A., A. Ranasinghe, and S. B. Weisberg. 2009. “Assessing Ecological Integrity in Marine Waters, Using Multiple Indices and Ecosystem Components: Challenges for the Future.” *Marine Pollution Bulletin* 59 (1-3):1–4.

Cao, Y., and C. Hawkins. 2011. “The Comparability of Bioassessments: A Review of Conceptual and Methodological Issues.” *Journal of the North American Benthological Society* 30 (3):680–701. <https://doi.org/10.1899/10-067.1>.

Chang, Winston, Joe Cheng, JJ Allaire, Yihui Xie, and Jonathan McPherson. 2018. *Shiny: Web Application Framework for R*. <https://CRAN.R-project.org/package=shiny>.

Cheng, Joe, Bhaskar Karambelkar, and Yihui Xie. 2018. *Leaflet: Create Interactive Web Maps with the Javascript ’Leaflet’ Library*. <https://CRAN.R-project.org/package=leaflet>.

Collins, S. L. 2016. “Vegetation Science in the Age of Big Data.” *Journal of Vegetation Science* 27 (5):865–67. <https://doi.org/10.1111/jvs.12459>.

Dale, V. H., and S. C. Beyeler. 2001. “Challenges in the Development and Use of Ecological Indicators.” *Ecological Indicators* 1 (1):3–10. <https://doi.org/10.1016/S1470-160X(01)00003-6>.

De Caceres, Miquel, and Pierre Legendre. 2009. *Associations Between Species and Groups of Sites: Indices and Statistical Inference*. *Ecology*. <http://sites.google.com/site/miqueldecaceres/>.

Fore, L. S., and C. Grafe. 2002. “Using Diatoms to Assess the Biological Condition of Large Rivers in Idaho (U.S.A.).” *Freshwater Biology* 47 (10):2015–37. <https://doi.org/10.1046/j.1365-2427.2002.00948.x>.

Grand, A., C. Wilkinson, K. Bultitude, and A. F. T. Winfield. 2012. “Open Science: A New "Trust Technology"?” *Science Communication* 34 (5):679–89. <https://doi.org/10.1177%2F1075547012443021>.

Hampton, S. E., S. S. Anderson, S. C. Bagby, C. Gries, X. Han, E. M. Hart, M. B. Jones, et al. 2015. “The Tao of Open Science for Ecology.” *Ecosphere* 6 (7):1–13. <https://doi.org/10.1890/ES14-00402.1>.

Hampton, S. E., C. A. Strasser, J. J. Tewksbury, W. K. Gram, A. E. Budden, A. L. Batcheller, C. S. Duke, and J. H. Porter. 2016. “Big Data and the Future of Ecology.” *Frontiers in Ecology and the Environment* 11 (3):156–62. <https://doi.org/10.1890/120103>.

Hering, D., A. Borja, J. Carstensen, L. Carvalho, M. Elliott, C. K. Field, A. S. Heiskanen, et al. 2010. “The European Water Framework Directive at the Age of 10: A Critical Review of the Achievements with Recommendations for the Future.” *Science of the Total Environment* 408 (19):4007–19. <https://doi.org/10.1016/j.scitotenv.2010.05.031>.

Howard, J. K., K. A. Fesenmyer, T. E Grantham, J. H. Viers, P. R. Ode, P. B. Moyle, S. J. Kupferburg, et al. 2018. “A Freshwater Conservation Blueprint for California: Prioritizing Watersheds for Freshwater Biodiversity.” *Freshwater Science* 37 (2):417–31. <https://doi.org/10.1086/697996>.

Ihle, M., I. S. Winney, A. Krystalli, and M. Croucher. 2017. “Striving for Transparent and Credible Research: Practical Guidelines for Behavioral Ecologists.” *Behavioral Ecology* 28 (2):348–54. <https://doi.org/10.1093/beheco/arx003>.

Jackson, S., and W. Davis. 1994. “Meeting the Goal of Biological Integrity in Water-Resource Programs in the US Environmental Protection Agency.” *Journal of the North American Benthological Society* 13 (4):592–97. <https://doi.org/10.2307/1467854>.

Karr, J. R., K. D. Fausch, P. L. Angermeier, P. R. Yant, and I. J. Schlosser. 1986. “Assessing Biological Integrity in Running Waters: A Method and Its Rationale.” Special Publication 5. Champaign, Illinois: Illinois Natural History Survey.

Kelly, M. G., S. Birk, N. J. Willby, L. Denys, S.Drakare, M. Kahlert, S. M. Karjalainen, et al. 2016. “Redundancy in the Ecological Assessment of Lakes: Are Phytoplankton, Macrophytes and Phytobenthos All Necessary?” *Science of the Total Environment* 15 (568):594–602. <https://doi.org/10.1016/j.scitotenv.2016.02.024>.

Kerans, B. L., and J. R. Karr. 1994. “A Benthic Index of Biotic Integrity (B-IBI) for Rivers of the Tennessee Valley.” *Ecological Applications* 4 (4):768–85. <https://doi.org/10.2307/1942007>.

Kuehne, L. M., J. D. Olden, A. L. Strecker, J. J. Lawler, and D. M. Theobald. 2017. “Past, Present, and Future of Ecological Integrity Assessment for Fresh Waters.” *Frontiers in Ecology and the Environment* 15 (4):197–205. <https://doi.org/10.1002/fee.1483>.

Liaw, Andy, and Matthew Wiener. 2002. “Classification and Regression by randomForest.” *R News* 2 (3):18–22. <http://CRAN.R-project.org/doc/Rnews/>.

Linke, S., E. Turak, and J. Nel. 2011. “Freshwater Conservation Planning: The Case for Systematic Approaches.” *Freshwater Biology* 56 (1):6–20. <https://doi.org/10.1111/j.1365-2427.2010.02456.x>.

Lortie, C. J. 2014. “Formalized Synthesis Opportunities for Ecology: Systematic Reviews and Meta-Analyses.” *OIKOS* 123 (8):897–902. <https://doi.org/10.1111/j.1600-0706.2013.00970.x>.

Lowndes, J. S. S., B. D. Best, C. Scarborough, J. C. Afflerbach, M. R. Frazier, C. C. O’Hara, N. Jiang, and B. S. Halpern. 2017. “Our Path to Better Science in Less Time Using Open Data Science Tools.” *Nature Ecology & Evolution* 1 (0160):1–7. <https://doi.org/10.1038/s41559-017-0160>.

Marwick, Ben, Jade d’Alpoim Guedes, C. Michael Barton, Lynsey A. Bates, Michael Baxter, Andrew Bevan, Elizabeth A. Bollwerk, et al. 2016. “Open Science in Archaeology.” *SAA Archaeological Record* 17 (4):8–14. <https://doi.org/10.17605/OSF.IO/3D6XX>.

McDonald, M., R. L. Blair, D. Bolgrien, B. S. Brown, J. J. Dlugosz, S. S. Hale, S. F. Hedtke, et al. 2004. “The US Environmental Protection Agency’s Environmental Monitoring and Assessment Program.” In *Environmental Monitoring*, edited by G. B. Wiersma, 649–68. Boca Raton, Florida, USA: CRC Press.

Michener, W. K., J. W. Brunt, J. J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. “Nongeospatial Metadata for the Ecological Sciences.” *Ecological Applications* 7 (1):330–42.

Nichols, S. J., L. A. Barmuta, B. C. Chessman, P. E. Davies, F. J. Dyer, E. T. Harrison, C. P. Hawkins, et al. 2016. “The Imperative Need for Nationally Coordinated Bioassessment of Rivers and Streams.” *Marine and Freshwater Research* 68 (4):599–613. <https://doi.org/10.1071/MF15329>.

Niemi, G. J., and M. E. McDonald. 2004. “Application of Ecological Indicators.” *Annual Review of Ecology, Evolution, and Systematics* 35:89–111.

Oksanen, Jari, F. Guillaume Blanchet, Michael Friendly, Roeland Kindt, Pierre Legendre, Dan McGlinn, Peter R. Minchin, et al. 2018. *Vegan: Community Ecology Package*. <https://CRAN.R-project.org/package=vegan>.

Pebesma, Edzer. 2018. *Sf: Simple Features for R*. <https://CRAN.R-project.org/package=sf>.

Pinheiro, Jose, Douglas Bates, Saikat DebRoy, Deepayan Sarkar, and R Core Team. 2018. *nlme: Linear and Nonlinear Mixed Effects Models*. <https://CRAN.R-project.org/package=nlme>.

Poikane, S., N. Zampoukas, A. Borja, S. P. Davies, W. van de Bund, and S. Birk. 2014. “Intercalibration of Aquatic Ecological Assessment Methods in the European Union.” *Environmental Science & Policy* 44:237–46. <https://doi.org/10.1016/j.envsci.2014.08.006>.

RDCT (R Development Core Team). 2018. “R: A language and environment for statistical computing, v3.5.1. R Foundation for Statistical Computing, Vienna, Austria.”

Reynoldson, T. B., R. H. Norris, V. H. Resh, K. E. Day, and D. M. Rosenberg. 1997. “The Reference Condition: A Comparison of Multimetric and Multivariate Approaches to Assess Water-Quality Impairment Using Benthic Macroinvertebrates.” *Journal of the North American Benthological Society* 16 (4):833–52. <https://doi.org/10.2307/1468175>.

Stein, E. D., M. Brinson, M. C. Rains, W. Kleindl, and F. R. Hauer. 2009. “Wetland Assessment Alphabet Soup: How to Choose (or Not Choose) the Right Assessment Method.” *Society of Wetland Scientists Bulletin* 26:20–24.

Stoddard, J. L., A. T. Herlihy, D. V. Peck, R. M. Hughes, T. R. Whittier, and E. Tarquinio. 2008. “A Process for Creating Multimetric Indices for Large-Scale Aquatic Surveys.” *Journal of the North American Benthological Society* 27 (4):878–91. <https://doi.org/10.1899/08-053.1>.

Stoddard, J. L., D. P. Larsen, C. P. Hawkins, R. K. Johnson, and R. H. Norris. 2006. “Setting Expectations for the Ecological Condition of Streams: The Concept of Reference Condition.” *Ecological Applications* 16 (4):1267–76. <https://doi.org/10.1890/1051-0761(2006)016[1267:SEFTEC]2.0.CO;2>.

Wickham, H. 2014. “Tidy Data.” *Journal of Statistical Software* 59 (10):1–23. <https://doi.org/10.18637/jss.v059.i10>.

———. 2015. *R Packages*. Sebastopol, California: O’Reilly.

Wickham, Hadley. 2009. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag. New York. <http://had.co.nz/ggplot2/book>.

———. 2017. *Tidyverse: Easily Install and Load the ’Tidyverse’*. <https://CRAN.R-project.org/package=tidyverse>.

Wickham, Hadley, Jim Hester, and Winston Chang. 2018. *Devtools: Tools to Make Developing R Packages Easier*. <https://CRAN.R-project.org/package=devtools>.

Wilkinson, L. 2005. *The Grammar of Graphics*. Second. New York: Statistics; Computing, Springer.

Wood, S. N. 2017. *Generalized Additive Models: An Introduction with R*. 2nd ed. London, United Kingdom: Chapman; Hall, CRC Press.

Xie, Yihui. 2015. *Dynamic Documents with R and Knitr*. 2nd ed. Boca Raton, Florida: Chapman; Hall/CRC. <https://yihui.name/knitr/>.