Open-Source Collaborations for a Sustainable Water Future

Policy Brief | September 2024

POLICY RECOMMENDATIONS

To ensure the effectiveness and sustainability of water data tools, we recommend that organizations involved in water data tool development:

- Implement user-driven design to encourage broad adoption.
- Use open-source development processes to reduce costs and increase durability.
- Prioritize funding for maintenance to ensure sustained impact.

WHAT IS USER-DRIVEN DESIGN?

A design approach that actively involves end-users throughout the planning, development, and testing stages to ensure the final product meets their needs and preferences.

WHAT IS OPEN-SOURCE DEVELOPMENT?

A collaborative development method where the source code is freely available for anyone to view, modify, and enhance, fostering community contribution and transparency.

INTRODUCTION

Water monitoring programs in the United States generate hundreds of millions of data points that could inform critical decisions made by water resource managers, regulatory agencies, and water suppliers. Despite the hundreds of millions of dollars spent each year to generate data, far fewer investments have been made to ensure that these data are used in decision-making processes (USGS 1995). Comparatively modest investments in automated data extraction and harmonization tools could greatly increase the value of these already collected and stored data.

BARRIERS TO INFORMED DECISION-MAKING

Water resource managers, regulators, and regulated entities depend on water data to comply with regulations like the Clean Water Act and make

ECOSYSTEM SCIENCE AND SUSTAINABILITY
COLORADO STATE UNIVERSITY

RADICAL OPEN SCIENCE SYNDICATE

Internet of Water COALITION

decisions related to water supply, public safety, environmental health, and more. The data used for these decisions is often sparse, difficult to access, and of varying quality, making it challenging for decision-makers to distill actionable insights. The process of making use of such diffuse data demands both technical proficiency and deep subject-matter knowledge, resulting in a significant bottleneck for effective water resource management. Despite the rising demand for this combined expertise, high technical barriers remain to accessing the vast catalog of water data available from local, state, federal, and private organizations.

Data Deficit
Gaps in data collection and sharing make it difficult for managers to make fully informed decisions. Even when data has been collected and shared, it is often challenging to find or access.

Lack of Transparency
Even when data is findable and accessible, important details such as who collected it, when, and how, often remain unclear. This information is critical to ensure appropriate use of the data.

Data Inaccessibility
Better tools are needed to extract, analyze, and harmonize data from different sources and to create consistent workflows.

User Interface Challenges
Many existing data systems do not provide user-friendly data retrieval, processing, and visualization and lack crucial features like geospatial functions and cross-platform integration, impeding efficient analysis and decision-making.

These common challenges often prevent water resource managers from using the most relevant data for their decisions. Moreover, employing specialists to navigate the complex network of existing data systems costs managers significant time and money.

THE CHALLENGE OF SUSTAINABLE WATER DATA ANALYTICS TOOLS

There is a recurring pattern in the domain of water analytics: tools, packages, and interactive platforms are developed, utilized briefly during grant periods, and subsequently abandoned. This cycle of creation, deterioration, and abandonment erodes trust in open-source development and leads to redundant efforts across the community.

Some larger agencies, private organizations, and academic institutions have invested in open-source tools that lower the technical barriers for subject-matter experts to access diverse water datasets. These tools–often built using open-source programming languages like R or Python–enable easier downloading, cleaning, and analysis of water data, or provide interactive web applications that eliminate the need for coding. While these tools can be simple and powerful, reducing the cost and time it takes to make data-informed decisions, they can also be brittle, error-prone, and too specific to the users who built them. In addition, many of these tools have low visibility, leading to redundant efforts, especially when developed in closed-source environments. Resilient, long-lasting, and continuously refined tools like the USGS-maintained package "dataRetrieval" are the exception to the general rule.

Even the most successful tools risk obsolescence or abandonment due to the ongoing effort required for maintenance and dependence on external projects that may undergo sudden changes or updates. Proposing the development of a new tool will almost always be more attractive to potential funders (foundations, federal agencies, or otherwise) than requesting funding for the ongoing maintenance of an existing tool. This mismatch between investment, which often prioritizes new tools, and the need, which is primarily focused on the maintenance and growth of existing tools, is common in software development.

To account for the missed costs of maintenance, software developers have created an approach that captures the full development cost of new tools called the Total Cost of Ownership. This approach encourages operators to budget for both new development and maintenance costs over a tool's lifecycle, typically on decadal timescales (Sneed 2004). The primary takeaway is that maintenance is often at least as costly as development and frequently much more expensive over longer time scales.

We estimate the combined costs of development and maintenance of water data tools to range from approximately $300,000 to $500,000 per year. Funding at this level would support three full-time employees filling the following roles:

- Outreach Specialist: collects user feedback on tool design, surveys users as the tool is developed, and ensures broad contributions to the code base.
- Data Scientist: leads open source tool development, tests tool performance, and integrates user feedback.
- Project Manager: Oversees documentation and ensures alignment and adaptation with evolving end-user needs. (While this person is not the lead tool developer, it is recommended that they have a strong familiarity with data science).

POLICY RECOMMENDATIONS

To ensure the effectiveness and sustainability of data tools, we recommend that organizations involved in tool development:

- **Implement user-driven design to encourage broad adoption.** To create more robust tools that broaden the use of public water data and associated tools, public agencies and funders should prioritize a user-driven design process, incorporating input from end-users throughout tool planning, development, and maintenance. User engagement can include informal interviews, surveys, and user experience testing. User testing should be included from the beginning of any project and maintained throughout the tool development lifecycle, with clear guidance for how end-users can provide feedback.
- **Utilize an open-source development process to reduce costs and increase durability.** While there exists a substantial community of technical professionals capable of developing sophisticated tools for public water data accessibility, closed-source development environments frequently result in isolated and redundant solutions. In contrast, well-managed open-source tools with clear contribution guidelines and defined goals facilitate the consolidation of distributed expertise into unified platforms. Such approaches yield robust applications characterized by broad functionality and a diverse user and contributor base.
- **Prioritize funding for maintenance to ensure sustained impact.** It is critical to account for both development and long-term maintenance costs when developing funding and budget forecasts. Maintenance expenses are typically estimated to be at least as much as initial development costs. Therefore, organizations developing water data tools should approach budgeting with long-term maintenance as the primary fiscal constraint, rather than initial development costs (Sneed 2004).

CASE STUDY: EPA & ROSS COLLABORATION

With support from the Internet of Water, the Radical Open Science Syndicate (ROSS) at Colorado State University collaborated with the Environmental Protection Agency (EPA) to develop the Tools for Automated Data Analysis (TADA). TADA aims to simplify data processing for the over 300 million water quality observations collected in the United States over the past 70 years (updated from Read et al., 2017).

The EPA is leading the development of TADA, with ROSS as a key collaborator contributing critical functionality. This partnership arose from the independent recognition by the EPA, the Internet of Water Coalition (IoW), and ROSS of a vital need to streamline access to public water quality data for state and local regulators and regulated entities through the Water Quality Portal (WQP). The WQP contains millions of water quality measurements gathered by more than 400 state, local, federal, and nonprofit entities (Read et al. 2017). While these data are accessible through automated pipelines (APIs), end-users often encounter numerous challenges related to data selection, harmonization across diverse sources, and determining suitability for specific use cases.

Each of our organizations (the EPA, ROSS, and IoW) had independently proposed the development of pipelines to facilitate the use of water quality data. Upon securing funding from IoW, ROSS opted to collaborate with the EPA rather than create a separate tool. Given our aligned objectives, we swiftly identified areas for improvement and defined how our team could contribute to TADA, specifically by adding geospatial workflows identified as a need that the EPA lacked the capacity to develop. This collaboration has proven highly successful, with ROSS enhancing functionality and modifying workflows that enable end-users to easily compare data and extract large datasets. The successful integration of our teams was the result of several key components:

- **User-driven Design:** Both the EPA TADA team and the ROSS group began our projects by engaging the water data user community, which includes state, local, tribal, and other government agencies that regulate water quality or are responsible for meeting those regulations. In developing EPA TADA, our team and the EPA surveyed over 100 stakeholders to inform tool design and functionality, greatly increasing its utility and use. This initial step is likely the most important for ensuring overall project quality, helping to build an engaged user community that will support and advocate for further tool development and maintenance.
- **A Well-Managed and Clear Open-Source Development Process:** A key factor in our success was the EPA's earlier decision to develop TADA using an open-source approach. This approach allowed us to examine the project's scope and ambition by accessing their codebase before contacting the EPA. This transparency enabled ROSS to understand the project's goals and current progress, identify areas where ROSS could contribute effectively, and align expertise with the project's needs efficiently. In addition, EPA's use of open-source software best practices throughout the development lifecycle has enabled our team to identify problems with the existing tool and propose and implement our own solutions.

- **Combined Funding from Diverse Sources:** Our collaboration with the EPA TADA team exemplifies the power of leveraging diverse funding sources in an open-source environment. Working collaboratively with the EPA, ROSS was able to implement many of the capabilities the EPA envisioned but did not have the capacity to develop. This collaborative approach resulted in a more comprehensive tool, delivering enhanced value and functionality that surpassed what either party could have achieved independently with their limited individual resources. This diversified funding helped TADA in its initial development, and it should be a continued priority for funding the maintenance of the tool. Dozens of organizations are stakeholders and end-users of the TADA tool, and future maintenance costs and efforts could be contributed from this broad user base and additional private or public grant dollars.

References

Intergovernmental Task Force on Monitoring Water Quality (US). (1995). The Strategy for Improving Water-quality Monitoring in the United States: Final Report of the Intergovernmental Task Force on Monitoring Water Quality (Vol. 2). Task Force.

Read, E. K., Carr, L., De Cicco, L., Dugan, H. A., Hanson, P. C., Hart, J. A., ... & Winslow, L. A. (2017). Water quality data for national-scale aquatic research: The Water Quality Portal. Water Resources Research, 53(2), 1735-1745.

Sneed, H. M. (2004, September). A cost model for software maintenance & evolution. In 20th IEEE International Conference on Software Maintenance, 2004. Proceedings. (pp. 264-273). IEEE.

ECOSYSTEM SCIENCE AND SUSTAINABILITY
COLORADO STATE UNIVERSITY

RADICAL OPEN SCIENCE SYNDICATE

Internet of Water COALITION