What the drive for open science data can learn from the evolving history of open government data

March 18 2021, by Stefaan G. Verhulst, Andrew J. Zahuranec and Andrew Young

Credit: Kaique/Pexels

Nineteen years ago, a group of international researchers met in Budapest to discuss a persistent problem. While experts published an enormous amount of scientific and scholarly material, few of these works were accessible. New research remained locked behind paywalls run by academic journals. The result was researchers struggled to learn from

one another. They could not build on one another's findings to achieve new insights. In response to these problems, the group developed the [Budapest Open Access Initiative](link), a declaration calling for free and unrestricted access to scholarly journal literature in all academic fields.

In the years since, [open access](link) has become a priority for a growing number of [universities](link), [governments](link), and [journals](link). But while access to scientific literature has increased, access to the [scientific data](link) underlying this research remains extremely limited. Researchers can increasingly see what their colleagues are doing but, in an era defined by the [replication crisis](link), they cannot access the data to reproduce the findings or analyze it to produce new findings. In some cases there are good reasons to keep access to the data limited—such as confidentiality or sensitivity concerns—yet in many other cases data hoarding still reigns.

To make scientific research data open to citizens and scientists alike, open science data advocates can learn from open data efforts in other domains. By looking at the evolving history of the open government data movement, scientists can see both limitations to [current approaches](link) and identify ways to move forward from them.

The three waves of open government data

The term open data did not [appear](link) until 1995, but the movement to open government data has a longer history. With roots in mid-century freedom of information legislation, open data emerged first as an attempt to push the boundaries of transparency and accessibility. This approach, part of the "first wave of open data" centered on removing secrecy in response to specific inquiries. While it had value it was also limited. It primarily benefited journalists, lawyers, and activists—those with the time, resources, and expertise needed to regularly query the government with specific requests.

As the Internet moved into the [Web 2.0 era](#) of the 2000s, new approaches began to emerge. Open government data came to be seen not just as a way to ensure accountability but a way to improve the processes of government itself. This second wave of open data prioritized problem solving.

It was, as former deputy chief technology officer for open government Beth Noveck once [noted](#), "not about transparent government. Simply throwing data over the transom doesn't change how government works. It doesn't get anybody to do anything with that data to change lives, to solve problems, and it doesn't change government. […] It's not even producing accountability as well as it might if we took the next step of combining participation and collaboration with transparency to transform how we work."

This conception allowed open government data to benefit a broader cohort, one that included civic technologists, governments, corporations, and start-ups. New York activists, for instance, [used](#) open government datasets to reveal improper ticketing practices by the New York Police Department. In Brazil, open datasets [supported](#) critical anti-corruption work. In Ghana, open data [helped](#) smallholder farmers sell their crops at better prices.

Yet, this approach too had its limitations. Data could be released without a clear sense of how others would use it, leading to a large number of datasets about issues unimportant to the public. Meanwhile, the desire to release datasets often led to a focus on assets that already existed, which benefited large institutions (often national governments) over smaller ones with less resources (such as local governments).

Recognizing these limitations, [a third wave of open data](#) has begun to emerge. This wave involves data holders across sectors and regions adopting a purpose-driven approach to making data accessible to the

benefit of community-based organizations, NGOs, academics, and small businesses. It seeks not just to open data for the sake of opening but to use collaborations to re-use assets that will be impactful. By paying as much attention to the demand for data as the supply, it concerns itself with the broader context within which data is produced and consumed. What's more, it asks how data held by businesses and other stakeholders can supplement those assets held by governments through data collaboratives..

This approach to data is still emerging but can be seen in many of the responses to COVID-19, which have relied heavily on collaborative methods. Efforts such as the NYC Recovery Data Partnership have combined public and private data assets on a local level to address public needs.

Much as the open government data movement has come to accept demand-driven, collaborative methods, so too can the movement to open science data. By recognizing the value of responding to more than just the "usual suspects," advocates can open the possibility of new and innovative research on pressing issues in their fields. They can allow professionals in other fields and domains the chance to build on their research.

Where to start

Advocates of open research data can learn from the open government data movement. As the last three decades show, opening data requires credible action on behalf of researchers, data providers, and intermediaries to foster a data ecosystem that sustains and nurtures collaboration. It requires organizations to make real commitments to openness.

While these efforts will not be easy, our research at The GovLab points

to several actions that organizations can take to foster an ecosystem that encourages openness. As we argue in our recent report, [The "Third Wave of Open Data](#)," open data can flourish if organizations focus collectively on:

Fostering and Distributing Institutional Data Capacity: In the public sector, data science capacity has often been relegated to small teams within organizations. This tendency has meant that attempts to use data are often ad hoc and isolated, with work siloed according to the field or skills. Much like those government organizations that have embraced the third wave, open research data advocates can try to build a culture of learning in their institutions, encouraging professional development and training programs that help low-ranking and senior researchers alike (re)use data in their daily operations.

Articulating Value and Building an Impact Evidence Base: In earlier waves of the open government data movement, advocates often started and ended their calls for open data by emphasizing the norms of transparency or accountability. While not wrong, these arguments tended to be less compelling to officials and members of the public who wanted to understand the tangible ways open data would improve their lives. Learning from this experience, open science data advocates can compile clear and specific uses for open science data to demonstrate the ways it strengthens existing research methods.

Creating New Data Intermediaries: Collaborating with outside organizations can be costly in terms of time, resources, and labor. Organizations like [Open North](#), [BrightHive](#), and StatsNZ's [Data Ventures](#) have emerged to address these costs. These organizations help public-facing organizations engage with possible collaborators by ensuring data is interoperable, providing mechanisms to securely share assets, and building trust between parties. Similar organizations could prove useful for sharing open science data.

Establishing Governance Frameworks and Seeking Regulatory Clarity: A recent MIT survey [found](#) that 64% of business executives in the United States are reluctant to embrace open data because of regulatory uncertainty. This statistic is telling Though a lack of regulation is often seen as offering organizations flexibility, the decades-long failure to develop policies around data reuse has instead disincentivized sharing. Just as organizations like the European Union have recently tried to [develop](#) strategies to organize the reuse of public and private data stores, open science data advocates might develop similar policies, plans, and procedures laying out expectations concerning the data they use.

Creating Technical Infrastructure for Reuse: In many countries, open government datasets are facilitated by open data portals. Sites like [data.gov](#) commingle various institutional datasets and allow users to browse, filter, search, and download data to their machines. While this approach can be used for open science data, additional technical infrastructure is likely for data users and suppliers to improve institutional capacity. As John Wilbanks of Sage Bionetworks [has argued](#), institutions might explore ways to subsidize computing capacity among target users and demographics, especially in fields where the datasets in question are prohibitively large and complex.

Fostering Public Data Competence: Open government data advocates in places like Taiwan have [sought to encourage](#) data competence among the public. These leaders argue that anyone should be able to participate fully in data projects, not just as consumers but producers of data-driven solutions that can improve their lives. To encourage public involvement in scientific research and foster new and innovative applications of data, open science data advocates might seek out ways to engage with the public, such as through research challenges and [participatory agenda-setting](#).

Tracking, Monitoring, and Clarifying Decision and Data Provenance: Decision provenance [entails](#) identifying decision points impacting data's collection, processing, sharing, analysis, and (re)use to determine which parties influence it. As third wave advocates have come to understand, awareness of these decision points is crucial in proactively identifying gaps and biases, both of which can undermine project goals. As such, open science data practitioners might create processes that allow others to understand the context from which data emerged and the harms that might result from inadvertent use.

Creating and Empowering Data Stewards:Finally, a core aspect of open government data work amid the third wave has been acknowledgement that data sharing and collaboration needs leaders who can champion it. These data stewards—responsible data leaders empowered by organizations to identify opportunities for data sharing and seek new ways of creating public value—already [exist](#) in a number of public sector institutions, civil society organizations, and companies. To encourage the creation of similar leaders in research spaces, open science data advocates might create [training courses](#) and [professional networks](#) that encourage data stewardship skills.

Conclusion

Almost two decades since the Budapest Open Access Initiative, our understanding of science data has changed profoundly. As we enter a new decade, these changes can continue so long as they are shepharded by advocates committed to expanding access.

We at The GovLab encourage researchers, institutions, and others to look toward the example provided by the [open data](#) movement in [government](#) to transform the way they work. By learning from others and adopting their practices, open research [data](#) can be a possibility for any field.