Things Fall Apart: Lessons from a Defunded Data Repository

ESSAY

ALEX DE SHERBININ

ubiquity press

Chinua Achebe's award winning novel, *Things Fall Apart* (1958), chronicles the cultural disintegration that occurred during the early phases of the British colonization of Nigeria. Any moment of abrupt cultural change is unsettling, and its results are unpredictable. The same can be said of the current moment in which belief systems that were foundational in the new era of digital data preservation are being shaken to the core. Since the start of the second Trump administration, we have seen a progressive removal of climate, environmental, and socioeconomic data from federal websites, with further threats to dismantle data infrastructure looming.[1] This essay represents a short personal account of the data rescue efforts of the NASA Socioeconomic Data and Applications Center (SEDAC), which was run by Columbia University's Integrated Earth System Information (CIESIN) since the late 1990s, and for which I served as deputy manager and then manager for nearly 20 years. I share this account to offer lessons for other data repositories and to underscore the importance of solidarity in the data community in the face of shrinking resources and foundering political support.

SEDAC was established in the 1990s, as part of NASA's Mission to Planet Earth, to serve as a bridge between the Earth and social science research communities. It provided data complementary to NASA's satellite remote sensing assets—such as gridded data representing population distribution, economic activity, poverty and social vulnerability, as well as infrastructure and sustainability data sets—all with the goal of serving the human dimensions of global change research community. SEDAC's flagship data product, the Gridded Population of the World (GPW), was the first raster representation of global population, setting the stage for increasingly sophisticated modeled population distributions such as Landscan, WorldPop, and GHS-POP. The American Geophysical Union (AGU) recently recognized GPW as an 'impactful dataset,' defined as one that supports the broad spectrum of research, analysis, and decision-making. SEDAC data were also widely used by practitioners such as urban planners and humanitarian and development organizations, and were frequently used in policy-oriented reports.

For SEDAC, the rumblings in Washington DC began just days after the inauguration in late January 2025, when we were notified by NASA that terms related to environmental justice, diversity, equity, and inclusion needed to be scrubbed from all its data center websites. These sad but required changes became more impactful for SEDAC when, on March 7th, SEDAC received a stop work order from NASA, part of a larger set of orders that were issued to all Columbia University principal investigators of federally funded projects. Although concerning, we hoped that this could be reversed through negotiations, and indeed eventually, most of the federal funding was restored to Columbia as part of a settlement agreement. However, by late March, we learned that SEDAC was on a list of the Department of Government Efficiency (DOGE), which seemed to seal its fate. By early April, I checked the NASA Earthdata catalog and

1 For example, the Trump administration has proposed to dismantle the National Center for Atmospheric Research over alleged 'climate alarmism.'

CORRESPONDING AUTHOR:
Alex de Sherbinin, PhD
Center for Integrated Earth System Information (CIESIN), Columbia Climate School, US
amd155@columbia.edu

KEYWORDS:
data repositories; data preservation

TO CITE THIS ARTICLE:
de Sherbinin, A 2026 Things Fall Apart: Lessons from a Defunded Data Repository. *Data Science Journal*, 25: 9, pp. 1–5. DOI: https://doi.org/10.5334/dsj-2026-009

found that SEDAC data had been removed even though our servers were still running. At the time, NASA staff were instructed not to communicate with us, so we were in the dark as to the intent of this removal. The actual termination of the contract was confirmed by an official letter on April 30th, but our data rescue planning began in early April.

The disappearance of SEDAC data and the increasing number of third-party data producers and users who reached out to us asking why the DOIs no longer worked was a gut punch—perhaps more than the loss of the contract itself. The data, carefully curated and produced by CIESIN, were for many of us a labor of love to which we had dedicated years of our lives. To see that they were no longer available online was discouraging, to say the least. From that moment, we began exploring options for the long-term preservation of our data. In consultation with archival staff, we were determined to preserve the entire collection in one place. Across the US, we saw well-meaning efforts to download data and upload them in the cloud, with Google documents serving as pointers, but this scattershot approach meant that data discovery would be extremely difficult for future users.[2]

By way of background, SEDAC had 300 data sets at the time of the termination, comprising around 61,000 granules, for a total data volume of around 6 TB. SEDAC had close to 210,000 registered users on NASA Earthdata Login, and those users downloaded roughly 17,000 files per month. SEDAC data represented a mix of data sets developed by CIESIN under the NASA contract, as well as many data sets produced by third parties and curated by SEDAC because of their value to its user community. This relatively small but high-value collection of data has received many thousands of academic citations across various disciplines over the years (Downs, 2023). We were just beginning to systematically assess the broader impact of the data, including the degree to which SEDAC data were being used—often in conjunction with NASA remote sensing data—to make decisions in government, non-profits, and the private sector. SEDAC was certified as a trustworthy digital repository by the CoreTrustSeal Standards and Certification Board, and consistently received exceptional ratings in its Federal Contractor Performance Assessment Reports.

After their disappearance from Earthdata, our initial inclination was to offshore the data in a foreign repository. Two leaders in the Earth Science Information Partnership (ESIP), Ruth Duerr and Steven Diggs, met with us on Zoom, and identified CERN's Zenodo as a possible safe haven. We also held discussions with colleagues at Queens University in Ontario, which has a data collection in Borealis, a Canadian version of the Dataverse.[3] Both options were promising, but there were two obstacles. We were short-staffed and seeking to hire a programmer even before the event—and Zenodo required programming to implement—and we faced the prospect of looming layoffs, which meant that our data systems team of archivists and metadata specialists would not be able to work for more than a few weeks. Further, it seemed unreasonable to expect employees facing job loss to dedicate substantial time and effort on a data rescue effort when they rightfully needed to be thinking about finding new employment.[4]

Closer to home, we considered storage on servers at the Lamont-Doherty Earth Observatory, where CIESIN is located, but this would have required ongoing funding and maintenance. In accordance with our CoreTrustSeal data preservation plan,[5] which mentions an MOU with

2 As a stop gap, CIESIN also put up a Google Sheet on its website with pointers to the download links.

3 During internal brainstorming and discussions with Columbia colleagues, we also considered science organizations in Japan.

4 A total of 12 SEDAC staff lost their jobs; fortunately many were able to find new jobs quickly and a few took early retirement.

5 See SEDAC's CoreTrustSeal certification at https://doi.org/10.34894/9RCIVO, and particularly the response to Requirement 10 (R10: preservation plan) and the section on Insource/Outsource Partners, in which we stated 'The Libraries have a long-standing commitment to host SEDAC data resources should CIESIN's contract to operate SEDAC end. The Columbia University Libraries and CIESIN have executed a Memorandum of Understanding (MOU) in 2022 that has established CIESIN as a Partner of the Libraries for hosting works on the Academic Commons platform.' Note that under R10, we also cited the NASA Earth Science Data Preservation Content Specification, which states that 'data resulting from NASA's projects are a valuable resource that needs to be preserved for the benefit of future generations. In the near-term, as long as the projects' data are being used actively for scientific research, it continues to be important to provide easy access to data and services commensurate with current information technology. For the longer term, when the research community focus shifts toward new projects and observations, it is essential to preserve the previous project data and associated information… It is essential for NASA to preserve all the data and associated content beyond the lives of NASA's projects to meet NASA's near-term objective of providing access to data and services for active scientific research, as well as for long-term use, such as reprocessing with revised algorithms to support long-term continuity with new measurements and measurement techniques. Also, NASA has to ensure that the data and associated content are preserved and available at a time in the future when permanent archive agencies will assume responsibility.'

Columbia University Libraries for long-term data preservation, we also explored Columbia's Academic Commons, but file size limitations would have meant that we would have to split up the collection between larger and smaller data sets. Similarly, we discussed the possibility of placing our spatial data in Esri's Living Atlas (where many SEDAC data sets already have a home), but that would again mean splitting up the data, since SEDAC has a large holding of tabular data. Lastly, we held discussions with the Inter-University Consortium for Political and Social Research (ICPSR), but the subscription model of annual institutional fees would have been difficult to sustain for a small center like CIESIN dependent on declining soft money funding.

Finally, without much hope of a response, I submitted a query via the online support tool at the Harvard Dataverse (HDV), and was delighted when—just hours later—I received a response from Sonia Barbosa, Associate Director of Dataverse Support. Sonia connected us with Harvard's Jonathan Gilmour and Boston University's Kevin Lane of the new CAFE Climate and Health Research Coordinating Center Collection, a National Institutes of Health-funded joint BU-Harvard HDV collection.[6] Here was a solution that ticked all of our boxes. CAFE had funding to support staff who could ingest our data, work with us to extend their standard metadata schema to accommodate our custom fields, and instruct us on the use of HDV for future dissemination of data that were still in the pipeline. In addition, we were assured that HDV had its own endowment, ensuring a degree of sustainability and autonomy if (as happened) Harvard's federal funding was suspended. Finally, HDV is mirrored on Borealis, providing an additional layer of long-term preservation.

We began working with the CAFE staff in late April, sharing the Google Sheet of data sets with download links, and providing pointers to our metadata for bulk ingest. The CAFE team started dedicating serious effort to the data rescue effort by early June, and by early September, they had completed uploading more than two-thirds of our high-value data, with only our largest and oldest data sets remaining (work that is still in progress). By September 18th, we went live with a release of the SEDAC data collection.[7] Importantly, this would not have been possible without the dedicated work of the CIESIN data system team, who continued to care deeply about the stewardship of our data even after it became apparent that their jobs were on the line.

It was only as plans to transition our data were well under way that we discovered that NASA Earth Science Data and Information System (ESDIS) staff had worked quickly to scrape our data from our servers while they were still running, and had moved the data into the cloud. By June, following the contract termination, NASA staff could once again communicate with us, and we were asked to transfer DOIs to ESDIS via DataCite, which we were happy to do.[8] By June, SEDAC data were once again available for download via Earthdata, though some download links remain broken.[9] This is likely a result of the haste required to move the data to NASA's AWS cloud before SEDAC servers were shut down—a process that normally would have taken a number of months with careful QA/QC.

There are a number of lessons to be learned from this effort. For one, the data rescue would not have been possible had Columbia not retained copyright to the data. The university is highly reluctant to sign over intellectual property, and the series of events confirms the wisdom of this policy. A second lesson is the importance of generalist repositories to long-term data preservation. As a manager of a domain-specific repository, I had grown to respect the generalist repositories of this world (HDV, Zenodo, Figshare, GitHub) since they pushed us to innovate in certain areas, especially in regard to speeding up the archival process for third-party data providers. Yet, with SEDAC I felt a pride associated with serving the human dimensions community well. We regularly met with scientists and practitioners in our user

6 CAFE fortunately had received three years of funding in 2024. Note that Gilmour is no longer at Harvard but runs a nonprofit called The Impact Project devoted to data advocacy and demonstrating the importance of data for decision making.

7 See https://dataverse.harvard.edu/dataverse/SEDAC and a data search tool https://ciesin.columbia.edu/content/data.

8 Moving the data to HDV necessitated creating new DOIs for the same data sets. While less than ideal, we decided it was a small price to pay to ensure long-term access.

9 See https://www.earthdata.nasa.gov/news/data-from-sedac-available-again-earthdata-search.

community, we dedicated resources to data quality checks and complete data documentation that self-archiving repositories are unable to commit, and we packaged data in ways that were responsive to community needs. The upside of the generalists is the rapid publication of data tailored to academic data producers; the upside of SEDAC was a deep knowledge of our community through a user advisory group, engagement at conferences, hosting focused workshops, and specialized staff.[10] As Diggs (2025) notes, data may be 'rescued,' but without the specialized knowledge of domain-specific repository staff, we risk losing knowledge of the data themselves—for example, strengths and weaknesses among data sets and their fitness for use.

The last important lesson that should not be overlooked relates to internal files. CIESIN maintained a significant server infrastructure for both internal processing and external dissemination of data. These servers required expensive software licenses to manage and maintain security, a NASA requirement of all contractors, and when those licenses ended to save costs and various hardware components were shut down (per NASA requirements), we discovered system dependencies that meant that many of our input data and other internal files were no longer accessible. We moved too slowly to get data off these servers, even if only onto temporary solid-state drives, and much time and effort has gone into trying to 'liberate' those files in the messy aftermath of an abrupt shutdown. Fortunately, if our initial data rescue plans had failed, SEDAC maintained a full offline archive on physical media (DVDs and Blu-ray discs), and we were in the process of backing up that physical archive to AWS S3 Glacier Deep Archive. A last resort would have been to extract the data from the physical media and move them to a new active or dark archive, depending on the value of the data.

Cultural change is hard. As Diggs wrote for a recent National Academies workshop on Earth Observations and Data Stewardship (2025, p. 3):

> For thirty years, the data management playbook was straightforward: migrate datasets from short-lived project servers into federal long-term repositories. University research groups came and went. PIs retired. Grant funding dried up. Departmental servers failed. But federal agencies (NOAA's National Centers for Environmental Information, NASA's data centers, USGS archives) represented permanence. We built data management plans, repository certification standards, and community trust on this bedrock assumption.

These assumptions are proving to be tenuous, at best, given the anti-science predisposition of the current administration. While we can hope for better days ahead, irreparable damage has already been done. Data may be the 'new oil,' but not everyone sees it that way, and there need to be continual efforts to prove the societal benefits of data, together with well-managed and resourced repositories to preserve them. In the short-term, at least, solidarity and making do with less will be keys to survival for our community.[11]

ACKNOWLEDGEMENTS

SEDAC data systems team members Robert Downs, Merlie Hansen, Lisa Lukang, Joachim Schumacher, John Scialdone, and Sri Vinay deserve special commendation for their dedication to the data preservation efforts. In the pursuit of a new home for SEDAC data, ESIP Federation members Ruth Duerr and Steve Diggs were particularly helpful early interlocutors who provided important advice along the way. The Harvard Dataverse CAFE team—Sonia Barbosa, Jonathan Proctor (now of The Impact Project), Alexis Guanche, and particularly Emily Katz—deserve special recognition for their efforts to preserve SEDAC data. Finally, I would like to thank the many far-sighted civil servants at NASA for their support for SEDAC over the years.

10 We also frequently built specialized data visualization and analysis tools incorporating third party data.

11 For example, repositories may wish to develop mutual support agreements to aid one another in the unfortunate event of defunding to ensure that data are preserved. Consistent with networked resilience, it may be advisable to develop anticipatory agreements between repositories in different countries that cover specific research domains. The United States has supported data as a global public good for close to 50 years, and in this time of crisis, the governments and repositories of other countries may need to step up to avoid irretrievable data loss.

FUNDING INFORMATION

I would like to acknowledge funding from the Columbia University Research Stabilization Fund, supported by Columbia Global, for the project Strategic Pivot for CIESIN Post-SEDAC.

COMPETING INTERESTS

The author has no competing interests to declare.

AUTHOR INFORMATION

Dr. de Sherbinin served as deputy manager of the NASA SEDAC from 2008 to 2024 and as manager from 2024 to 2025. He also served as a member and then as chairperson of International Science Council's World Data System scientific committee from 2015 to 2021. He is currently the director of CIESIN and a senior research scientist at the Columbia Climate School where he remains engaged in research, data analysis and production, and teaching.

AUTHOR AFFILIATIONS

Alex de Sherbinin, PhD orcid.org/0000-0002-8875-4864
Center for Integrated Earth System Information (CIESIN), Columbia Climate School, US

REFERENCES

Achebe, C. (1958) *Things Fall Apart*. London, UK: Heinemann.

Diggs, S. (2025) 'An Ensemble Framework for Climate Data Preservation: Integrating Value, Vulnerability, and Observational Continuity', *White Paper prepared for the NASEM workshop on Future Directions for Earth Observations and Data Stewardship*.

Downs, R.R. (2023) 'Scientific Use and Impact of Socioeconomic Data across Disciplines', Prepared for Presentation to the *Research Data Alliance 21st Plenary (RDA P21), International Data Week, Salzburg, Austria*, 23–26 October 2023. Available at: https://doi.org/10.7916/szky-3127

TO CITE THIS ARTICLE:
de Sherbinin, A 2026 Things Fall Apart: Lessons from a Defunded Data Repository. *Data Science Journal*, 25: 9, pp. 1–5. DOI: https://doi.org/10.5334/dsj-2026-009

Submitted: 07 January 2026
Accepted: 13 February 2026
Published: 02 March 2026

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.