Open Science: Key to accelerating interdisciplinary research & education

*An **Earth Lab** white paper contributed to **CU Boulder's Academic Futures***

Authors: Jennifer Balch, Brian Johnson, Bill Travis, Leah Wasser, Maxwell Joseph, Megan Cattau, Melissa Maestas, and Chelsea Nagy

1. Why Open Science should be fundamental to our Academic Future

Knowledge is a global public good.[1] The new knowledge generated in our labs, libraries, field sites, and offices, belongs to all humankind.[2] And as most science is built on some level of public investment, knowledge generated by science, and the steps taken to produce that knowledge should be publicly available. Open science[3] is the process of making scientific methods and findings (data, software code, analysis workflows, and papers) accessible and reusable to others outside of the original research team. At Earth Lab we are committed to making the majority of our work open and reproducible.[4] We see **open science as a critical pathway to accelerating the best new environmental science from big data opportunities**.

Big data, and the big science that it enables, are inextricably connected to open science. Both are inherently complex, requiring large, interdisciplinary teams-from the Hadron collider, a coordinated effort of more than 10,000 physicists,[5] to the National Ecological Observatory Network, which represents a $400 million observing network involving hundreds of people. Open science is a key mechanism to make this happen, as **well-documented pieces of the scientific process can be shared and built upon by many**, facilitating large, multidisciplinary collaborations. Open science also improves the robustness of scientific findings and increases the rate of scientific discovery.[6] We see a future in which professional researchers, emerging student leaders, industry innovators, citizen scientists, and others contribute to the solutions to our most challenging environmental problems through an open science framework. The university becomes a node for incubation of ideas and collider spaces where academic, government and commercial partners meet physically and virtually to collaborate on scientific challenges. These nodes enable us to inform our research agendas and move research ideas and methods rapidly into practice to meet real world problems across sectors.

[1] Stiglitz, J.E. (1999). Knowledge as a global public good. In: Kaul, I., I Grunberg, and M. Stern (Eds.) *Global Public Goods: International Cooperation in the 21st Century* (pp. 308-325). Oxford University Press, New York.

[2] Callon, M. (1994). Is science a public good? Fifth Mullins Lecture, Virginia Polytechnic Institute, 23 March 1993. *Science, Technology, & Human Values* 19.4: 395-424.

[3] Nielsen, M. (2011). *Reinventing Discovery: The New Era of Networked Science*. Princeton University Press, New Jersey.

[4] Earth Lab. (2017). *Earth Lab Strategic Goals: 2018-2023*. https://docs.google.com/document/d/1XPFadac4xMMvfJWpA8w87hjfbSlpiKyD-tdkFIHcDyM/edit#

[5] Merali, Z. (2010). Physics: The large human collider. *Nature* 464: 482-484. http://www.nature.com/news/2010/100324/full/464482a.html

[6] Wicherts, J.M., M. Bakker, and D. Molenaar. (2011). Willingness to share research data Is related to the strength of the evidence and the quality of reporting of statistical results. *PLoS ONE* 6(11): e26828. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0026828.

Universities across the country recognize the opportunity in big data, yet do not focus on open science as key to unlocking the possibilities. Berkeley requires every incoming first-year student to take a core data science class. The University of Michigan is hiring 35 new faculty in big data. However, it is not just about a single class or needing new faculty. After skills in data science are learned **the next big step is committing to open science**.

There are fundamental systemic changes that are needed to accelerate interdisciplinary research and education with open science. Below we outline what those systemic changes are. CU Boulder has an opportunity to be a leader in this brave new world of open science.

2. Key strategic goal: The majority of CU Boulder's research will be open and reproducible by 2023.
What if CU Boulder made a commitment to making the majority of scientific research open and reproducible? These would be the important outcomes: i) open science would accelerate generation of new scientific insights; ii) anyone could contribute and learn, creating a global university that facilitates participation from students, scientists, and citizens ; iii) open science would help to integrate research and education; and iv) impact metrics would soar as a consequence of making different parts of the workflow usable and citable.

3. Benefits of Open Science

Open science leads to rapid discovery by providing scientific results, methods, and techniques that can be efficiently reproduced and built upon. The scientific process is made transparent by publishing or making publicly available data, code, and new tools as well as by having workflows that are packaged so that they can be easily shared and do not depend on the specific computer hardware set-up to run. This is a critical pathway for deriving the best, new environmental science,[7] reflected in NSF's approach to mandating data management plans, a reflection of the growing recognition of the importance of open, reproducible science. Below, we outline some key benefits of open science that provide a framework for specific best practices:

Open science directly benefits researchers

Open science can increase scientific rigor and impact. By ensuring that every result in a paper is reproducible, scientists build in a verification layer to their work, which quickly catches errors and provides a way for reviewers to better evaluate a manuscript. Data and code aside, open access publications are downloaded and cited at higher rates than closed access papers.[8] [9] In addition, open research is associated with increases in media citation, potential collaborators, job opportunities and

[7] Hampton S.E*., et al.* (2015). The Tao of open science for ecology. *Ecosphere* 6(7); Wolkovich E.M., J. Regetz, and M.I. O'Connor. (2012) Advances in global change research require open science by individual researchers. *Global Change Biology* 18(7): 2102-2110.

[8] Ottaviani, J. (2016). The post-embargo open access citation advantage: It exists (probably), its modest (usually), and the rich get richer (of course). *PLoS One* 11: 1–11.

[9] Tang, M., J.D. Bever, and F.H. Yu. (2017). Open access increases citations of papers in ecology. *Ecosphere* 8**,** 1–9.

funding opportunities.[10] Further, it is often easier to build upon previous scientific work when the data and code are made available, within individual research groups and for broader communities.

The data, code, and software are at least as valuable as a manuscript

Scientific manuscripts are only one component of the scholarship involved in a project. The data, code, and software used constitute building blocks for future work. All of these components can receive digital object identifiers (DOIs), through platforms such as Dryad (for data), GitHub and Zenodo (for code), and the Journal of Open Source Software (for software), the result being just a citable as a conventional publication. By releasing data, code, and software along with a paper with clear guidelines for use, scientists can increase the impact of their work.

Licenses enable reuse and modification, leading to rapid advances

Scientists may not realize that they can still have control over how their work is used, even if it is open. By specifying a license for code and/or data, authors can control who uses their research products, how they are permitted to use them, and how derivative works must be licensed (see https://choosealicense.com/). If something is released with no license, then by default no other parties can legally use, copy, distribute, or change the work that has been released, so it is not enough just to release code or data on the internet via a public link.

Research can be "closed" until the researcher releases it

To a large extent, science is a novelty economy: new ideas, new methods, new results, and new ways of learning about the world are highly valued, creating an incentive to publish first. For many scientists, this leads to fear of being "scooped", or having another group publish the same idea first. As a consequence, many scientists are reluctant to share new findings until papers are accepted or published, meaning they would prefer not to have a totally open workflow prior to publication. Research can be closed until authors publish, and still substantial benefits will come with making individual components of the workflow open at that time. Committing to open science requires a cultural shift and acceptance that a scientist's contribution will be valued and cited in additional ways.

4. Challenges

Fostering open science across the university will require **overcoming cultural, technical, and practical challenges** in how we conduct our research and in the reward structure.

Cultural Challenges

Culturally, the university will need to change our incentive structure. While publishing research findings is central to the academic reward system, publishing software and data is not as valued. The opportunity awaits for CU Boulder to lead the way in adopting incentives through changing the academic incentive structure to include rewarding making code and methods available. However, changing a long standing incentive process will be slow and will take time and effort. Open and reproducible science requires a time commitment to document code and keep the workflow organized. The promotion and tenure

[10] McKiernan et al. (2016). How open science helps researchers succeed. *eLife* 5:e16800. DOI: 10.7554/eLife.16800.

process and evaluation methods will need to be adjusted at every career level to acknowledge the value of the time spent making science open and reproducible.

Technical Challenges

Not all researchers have the technical skills needed to implement open science approaches. Further, learning the skills will take time. Incentives to publish code may encourage researchers to make the time to learn the tools required to document their workflows. But also, training is required to support faculty in the effort of learning new methods and approaches.[11] This will help ensure that open data and workflows are both usable and discoverable by the broader community. Earth Lab is committed to education of students, faculty and staff that supports open science workflows. Some of the technical barriers to open reproducible science have been reduced with the the advent of technology and platforms that make code easier to share and run on various computer systems, e.g., GitHub and Docker/Singularity.

Practical Challenges

The primary barrier for researchers to share code and data is often the time it takes to clean up their software, and document the data and work to prepare it for release.[12] Not all data and workflows can be shared, e.g., sensitive (human subjects data) or proprietary (industry specific) information. In these cases, however, it may be possible to share parts of the workflow - such as code or anonymized data. In situations where the code could be shared, but not the data, it may be helpful to other researchers to offer a proxy data set that is compatible with the code. Each research project working with sensitive or proprietary information will need to be evaluated individually. Another practical challenge is the cost of publication, as it is sometimes more expensive for researchers to publish their work in open access than closed, traditional journals. With broad acceptance of the newer, fully open access journals that do not charge authors at all,[13] the high costs of library subscriptions to closed-access journals could be reduced.

5. Key elements of the Earth Lab model

Open science can rapidly generate new insights when coupled with intentional mechanisms to facilitate synthesis and cross-pollination of ideas and approaches. Earth Lab fosters such a model that is built on promoting best practices for open science, enabling data- and compute-intensive work through tightly coupled analytics and computing support, educating students and scientists in how to do data-rich Earth Systems science, building spaces and events that facilitate authentic interaction, and fostering a postdoc community that is an energized network across campus.

Best practices are a pathway to capitalizing on big, diverse data, building multidisciplinary communities and accelerating data-intensive research. Earth Lab's best practices and guidelines for open,

[11] Hampton, S.E., et al. (2017). Skills and Knowledge for Data-Intensive Environmental Research. *BioScience* 67(6): 546-557. https://academic.oup.com/bioscience/article-abstract/67/6/546/3784601.

[12] LeVeque et al. (2012). Reproducible Research for Scientific Computing: Tools and Strategies for Changing the Culture. *Computing in Science & Engineering* 14(4): 13-17.

[13] Walt Crawford. (2014) 72% and 41%: A Gold OA 2011-2014 preview. https://walt.lishost.org/2015/08/72-and-41-a-gold-oa-2011-2014-preview/

reproducible science include: developing data management plans; encouraging use of R and python open source software languages; managing code development using version control tools like GitHub; capturing the computational environment and software dependencies using containerization software (e.g., Docker); automating the analysis workflow; and submitting data and code to public repositories. We are also developing shareable lessons and software tools to access and explore data and workflows to assist in review and reuse. And, we are training our students, postdocs, and faculty in these practices. These practices accomplish two objectives: i) making data open, accessible, and reusable and ii) making computation reproducible by recording the methods, protocols, and provenance of algorithms and data used in more detail than typically appears in a published journal article.

Tightly coupled analytics and computing support is critical for researchers and students. The data revolution, i.e., size, complexity, and diversity of data, is fundamentally challenging our traditional scientific approaches, and arguably we could not handle big data without open science. Data-driven hypothesis development, rapid theory testing, and constraints on classic statistical methods are all being challenged and reshaped. Methods for accessing, preparing and investigating complex, heterogeneous datasets are highly specialized and time consuming. Further hampering progress is the gap in science and engineering training in data science and big data technologies. Earth Lab has cultivated a strong multidisciplinary team with expertise in remote sensing, computing and visualization, open source software programming development, geospatial and temporal statistics, and machine learning algorithms to help researchers bring new data, analytics, and scalable computing to their science.

Data-intensive education builds data skills that become fully integrated in our courses. Targeted and community-driven trainings help up the game of current faculty and researchers on campus. Professional programs are critical for filling the major gap in training in data science for specific domains and increasing connections between the university and industry and federal partners. Earth Lab's education initiative is pathfinding new ways to provide online learning materials and create a rich environment for students and professional development through innovative, interdisciplinary course and curriculum design.

Spaces and events are idea incubators and facilitate cross pollination of ideas. Earth Lab provides communal working areas for postdocs, students, and affiliate faculty working on similar themes. This is a central element in promoting synthesis work. Earth Lab is such a space for Earth Systems science at CU Boulder. Through our data jams, cloud workshops, science slams, and public seminars we are holding events that help researchers and students share their core ideas and approaches to find opportunities for interdisciplinary research and education.

A cultivated postdoc community builds on open science. Postdocs are key players in facilitating interdisciplinary research at CU Boulder. Moreover, they are the engine that drives our research forward in leaps and bounds. Unlike a traditional postdoc experience, Earth Lab postdocs are co-located in a shared office space with other researchers from a broad range of areas of expertise. The postdocs work with PIs and collaborators from all over campus. This arrangement forms a node creating opportunities to interact that would not otherwise exist, allowing for a greater exchange of ideas across campus and opportunities to collaborate in interdisciplinary research. Flexibility in this stage of the postdoc career lends itself to building collaborations across campus. This flexibility is important for capitalizing on the collaboration potential of PIs in different fields with limited time but with strong potential overlap in research interests.

With a strong skill base and capacity for learning additional skills, postdocs are in a unique position to take advantage of the training opportunities offered through the Earth Lab, including training in best practices for open science. These skills make them more marketable and allow them to disseminate research to the broader community. Open science is more cost effective in the Earth Lab model because postdocs create reusable, extensible research products. Bundling scripts into packages and creating well-documented, accessible, user-friendly data products ensures that these products can continue to be used by Earth Lab and the broader community beyond the time that the postdoc is based at CU. This model is cost-effective and provides a means of retaining institutional knowledge. Further, the public availability of code packages, data products, and subsequent citations increases the visibility and reputation of CU.

6. What next?

Incentivizing and enabling open science at CU Boulder requires a university-wide strategic framework. This could include: 1) building data science into a core university curriculum, 2) encouraging adoption of a best-practices guide for open, reproducible science, following Earth Lab's lead, and 3) evaluating job and tenure candidates based on their contributions to open science. Existing groups on campus that are already committed to open science could support these efforts. The National Snow and Ice Data Center in CIRES curates large, complex satellite datasets, making these data discoverable, accessible, and consumable. Research Computing provides scalable data storage with the PetaLibrary, high-performance computing with Summit, and is working to bring cloud computing into the university. The Center for Research Data and Digital Scholarship, a partnership between the Libraries and Research Computing, is actively promoting and providing training in data management, digital scholarship, and the Open Science Framework. Last, Earth Lab is building upon these practices using open source software, making code and workflows available, developing and publishing online training material for everyone to use, and facilitating the networks that can rapidly build and deploy component pieces of open science.

The benefits of a commitment to make the majority of CU Boulder's research open and reproducible by 2023 would be tremendous. Open science would forge new pathways connecting together CU Boulder's strengths in earth, space, and social sciences, library sciences, advanced computing, and aerospace technologies. Capitalizing on existing infrastructure and capabilities, and creating new ways to educate and train students would position **CU Boulder as a national leader in big, open science**.