



## LEAP Data Management Plan

**Data Types Produced and Collected.** Six data types will be generated and collected across LEAP's lifespan, including:

1. **Scientific Data** produced during the course of LEAP's studies and research;
2. **Software and Models** trained on scientific observations of the Earth system;
3. **Curricula** related to the activities described in the Education Plan;
4. **Personnel Data** on the demographics, educational, and professional activities of LEAP personnel, collaborators and partners;
5. **Open Source Engagement Data** related to public, academic, and industry engagement with LEAP's research, website, and communication modules; and
6. **Assessment Data** gathered in the course of evaluating the Center itself.

**Scientific Data** will include experimental measurements, observational measurements, model simulations, data compiled and/ or processed from external sources, and metrics, toolkits, and scientific publications based upon data analyses. **Software and Models** will include ML algorithms and their pre-trained models. **Curricula** will include course outlines, lesson plans, reference materials, supporting software tools, curated data sets used for the undergraduate research courses, capstone project submissions, and presentations used across all education, training, and outreach activities. **Personnel Data** will consist of records for all LEAP personnel, collaborators, and partners' backgrounds, program activities, and achievements and, for the student and postdoctoral participants, subsequent employment. **Open Source and Engagement Data** will include information on open source disclosures, patents filed, patents awarded, license inquiries, and jobs created. **Evaluation Data** will include faculty, students, collaborators, and partners' responses to surveys, interviews, focus groups, course and program reviews, mentoring reports, data collected from collaboration platforms, communication exchanges, open archive sources (Google Scholar), and reports from LEAP's leadership.

**Data and Metadata Standards.** All LEAP participants will be required to document their data and software so that both can easily be distributed to others when appropriate. For physical data, documentation will include data source, data sampling, processing, transcription or digitization methods, experimental protocols, data scale and resolution, sampling design procedures, and instrument calibration methods, where applicable. For electronic data, documentation will include data source, file formats, types and structure, naming conventions, versioning and any validation procedures used to ensure data authenticity. Research software will be written in the language of the LEAP participants' choice. Publications will be preserved in PDF. Personnel and evaluation data will be kept in database, spreadsheet, and/ or PDF format, depending upon collection method and evaluator **Lang's** choice. Curricula and reports, including reports on open source disclosures, will be preserved in PDF. Policies for access and sharing and provisions for appropriate protection/ privacy are outlined below.

**Access and Sharing.** Distributable Scientific Data will be made available by academic license (see below) on LEAP's website, hosted by **Columbia**. Research publications will be freely and publicly accessible on the site as they become available, in versions that conform with publisher and copyright requirements. LEAP participants will be asked to make their Software and Models available from the website, if there is the potential for use by others. Curricula and Open Source and Engagement Data will also be made freely and publicly available by academic license on the website. Personnel Data will be available only to LEAP faculty, the Director's Council, and administrative staff support, and, in aggregate and anonymized form, to the NSF. Evaluation Data will be available to LEAP faculty, the Executive Committee's staff support, the External Advisory Board, and NSF in forms that conform with the Institutional Review Boards of all partner institutions.

**Provisions for Appropriate Protection of Privacy, Confidentiality, Security, Intellectual Property, or Other Rights or Requirements.** The major privacy and confidentiality issues that will be likely to arise

concern Scientific Data and Personnel Data. **Columbia's** Data Security standards will be applied to ensure that the latter are seen only by those authorized. Scientific Data will be made available (assuming that there are no restrictions on redistribution from outside data sources) within one year of the completion of the degree of the student who has collected the data, giving the student a chance to continue their research thread, unless another student is sharing the data, in which case that deadline will be extended to one year following that student's graduation. Similarly, with software. Research publications will be made freely and publicly available as they are published, assuming that there are no copyright issues. The curriculum and outreach material will be made freely and publicly within a year of the material being successfully implemented in a course, workshop, training session, or program.

**Policies and Provisions for Re-Use, Re-Distribution, and the Production Of Derivatives.** Free Software or Academic Free licenses will be required to download data from LEAP's website. Personnel Data will *not* be shared beyond LEAP faculty and staff, except in anonymized and/ or aggregated form with the NSF. Evaluation Data will be shared only with LEAP faculty, administrators, the Student Leadership Board, the Director's Council and External Advisory Board, and the NSF. Security for data on LEAP's website will be maintained by **Columbia** University Information Technology (CUIT), within the University's computer network, and behind the University's firewall. LEAP personnel activity data will be maintained by the LEAP staff support and as appropriate shared with faculty and evaluators. (This data is distinct from personnel records, which will be kept on secure servers maintained by LEAP's partner institutions, and will be accessible only by appropriate administrative staff at those institutions and made available to the LEAP project team only.)

**Plans for Archiving and Preservation of Access.** The long-term strategy for maintaining data access will be LEAP's website, which will be preserved by **Columbia** (the lead University) after STC funding ends, using **Columbia's Academic Commons** digital repository, which will be preserved by the University into the foreseeable future. Content in Academic Commons is freely available to the public. Center Director **Gentine** will work with LEAP leadership and staff on implementation of the data management plan, with the aid of **Columbia** University Information Technology (CUIT) and the Academic Commons staff. As noted above, project participants will provide documentation, data, and software for their research products, such that they will be distributable and usable for others. Documentation provided by student participants will be reviewed by the student's advisors before matriculation. Such data with documentation will then be available on an 'as is' basis; maintenance of any software will be the responsibility of the graduate student.