C-DEBI DATA MANAGEMENT PLAN

The C-DEBI data management plan (DMP) was developed to insure that C-DEBI data and products are archived, shared, and accessible for the long term. The data types and products covered by the C-DEBI DMP include a wide variety of geophysical, geological, geochemical, and biological information, in addition to education and outreach materials, technical documents, and samples. All data and information generated by STC-supported researchers as part of their C-DEBI projects is to be made publically available either following publication or within two (2) years of data generation. Any extension to this policy can only be made in consultation with the both C-DEBI Director and Associate Director.

C-DEBI Data Types

C-DEBI produces many kinds of products that will need long term archiving. These include diverse data sets (biological, chemical, physical, and geological), samples, peer-reviewed publications, technological advances with associated engineering drawings and software, educational/outreach materials (such as K-12 and community college lesson plans relating to subseafloor science), and model parameters (e.g., inputs, grids, reaction rates).

Biological products include, but are not limited to, molecular data, activity data (isotope abundance, community enzymatic, etc.), frozen samples, and living microbial strains, and post-processed molecular data (e.g., 16S rDNA and 16S rRNA sequences, single-cell genome, metagenome and metatranscriptome sequences). Non-biological data will include multi-beam maps, seismic reflection profiles, and thermal, chemical, and physical data from recovered samples of fluids, sediment, rocks, and experiments. Measurements also will be made *in situ* using borehole observatories, drilling platforms, cabled observatories, and coring facilities.

Phase 2 technological advances will include methods, software, and hardware, each needing different methods for archiving. For example, C-DEBI researchers will generate and evaluate (i) bioinformatic tools and scripts to process nucleic acid data; (ii) software for geophysical data acquisition, oceanographic instrument operation, and interactive data processing; and (iii) coupled fluid-energy-biochemical- microbial models of the subseafloor. All advances and the benchmarks we use to evaluate them will be hosted in publicly available repositories

Standards and Metadata

To ensure broad access, C-DEBI researchers will follow the metadata guides and references produced by the Marine Metadata Interoperability project (marinemetadata.org). All genomic sample collection will conform to the Genomic Standards Consortium's (GSC) minimum information about a genome sequence (MIGS) or the minimum information about a metagenome sequence (MIMS). These are the minimum standards adopted by the community; for C-DEBI, we might expand this list of metadata to include a broad range of additional data, depending on the type of sample collected (e.g., chemical concentrations, activity data, mineralogy, etc.).

C-DEBI researchers will follow the principles articulated by Goldstein et al. (2003, *Chemical Geology* 202, 1-4) regarding open reporting of analytical metadata. Calibration standards, standard reference materials, internal standards used to quantify analytical precision, and other items analyzed as part of good analytical practice will be fully described in relevant publications.

C-DEBI researchers will collect and process geophysical data using public domain tools (e.g., MBsystem, Seismic *nix) in standard formats, including metadata that describe acquisition parameters. C-DEBI researchers should collate sample and data locations and types following guidelines of the IODP (for drilling expeditions) and the Integrated Earth Data Applications office (for other expeditions).

Data Access and Data Sharing

All post-embargo data, including primary data, metadata, and derived data, will be deposited in appropriate internationally accessible data repositories and directly linked by the C-DP. The principal repository will be at the Biological and Chemical Oceanography Data Management Office (BCO-DMO). The C-DEBI DMI will work with BCO-DMO to make this location will either the primary host of the C-DEBI data or to have them provide stable links to data housed in other repositories (i.e., NCBI). Additionally, all genetic sequence data will be deposited into such databases as GenBank or the EBI European Nucleotide Archive. Data generated on IODP expeditions will be deposited in IODP databases. Most remaining data will be deposited in other publicly accessible archives, such as the Biological and Chemical Oceanography Data Management Office (BCO-DMO), the National Geophysical Data Center (NGDC), or the NSF-supported sediment chemistry database (SedDB), as appropriate. Expedition survey results, including underway data, will be submitted to the ship operator and appropriate national databases within a year of cruise completion.

Any product for which a suitable national repository does not exist, such as educational materials, outreach materials, and technical advances, will be posted directly in the C-DP. To the extent possible, all such products will also be described in peer-reviewed literature, to ensure public dissemination and long- term accessibility beyond C-DEBI.

All C-DEBI data and intellectual products (publications, technical advances, software, education and outreach materials) will be directly linkable through the C-DP. This portal will provide direct electronic access to these products and also link to a broad range of other C-DEBI-relevant data.

Physical materials from C-DEBI projects will be curated and archived at publicly accessible repositories and made available to outside investigators within a year of arrival at the repository. For example, cores from C-DEBI expeditions will be curated and archived at NSF-funded Rock and Core Repositories. Once microbial isolates have been obtained and described in appropriate journals, cultures will be deposited in publicly accessible collections, such as the American Type Culture Collection (ATCC) and the Deutsche Sammlung für Mikroorganismen und Zellkulturen (DSMZ). C-DEBI is also currently in discussion with Dr. Sherry Cady, who has taken charge of the Culture Collection of Microbes from Extreme Environments (CCMEE; Dick Castenholz's former collection of environmental samples and enrichments) for future culture archive efforts.

Any C-DEBI data or intellectual product will be owned by the team that generated it, but made publicly accessible online within two years of its creation (to allow time for publication). Every C-DEBI investigator will be personally responsible for transferring to C-DEBI (or an appropriate archive) all data associated with their project.

Prior to public access, any data and/or intellectual product will be maintained by the team that generated it and routinely backed up on external servers. For this data backup, C-DEBI and USC will provide sufficient server and hard-drive storage capabilities to members of the C-DEBI community, with suitable restrictions to keep each group's data private from other users.

Enforcing C-DEBI Data Management Requirements

Researchers, students and others receiving STC support will be provided with a copy of the C-DEBI data management plan, and required to certify that they have read the DMP and agree to comply. When annual reports are submitted, researchers, students and others receiving STC support are required to provide an explanation as to what data sets are housed in which repositories, and what data sets remain to be uploaded/archived. In addition, C-DEBI researchers will be required to provide keywords associated with the their data for linkage within the C-DEBI data portal. Failure to comply with these requirements will result in cessation of funding (if any remains to be awarded, for example during a second project year) and removal of access to C-DEBI resources (e.g., computer servers, bioinformatics help, small grant consideration, etc.) until compliance is achieved. Further, C-DEBI will report "out of compliance" events and the mitigating steps C-DEBI is taking to resolve the issue to the appropriate NSF Program officer.

Policies and Provisions for Re-Use, Re-Distribution, and Production of Derivatives

C-DEBI is actively establishing computational infrastructure to provide investigators and teams with access to appropriate hardware, software, and data for current scientific challenges, particularly microbial genomic research projects. This infrastructure includes server(s) to host and analyze project data, software licenses for researchers accessing the server(s), redundant hard drive space for data backup and network integration with the USC HPCC system, the C-DP (for hosting and linking derivative data products, metadata, and other intellectual products), and in-house bioinformaticians to advise and support C-DEBI investigators (Heidelberg and the Bioinformatics Specialist). Much of this in-house bioinformatics effort will focus on developing new means to analyze current data and producing high-quality programming code that will be available to all researchers, through open-source repositories like GitHub.

As described above, all C-DEBI data will be placed in publicly accessible databases. Further, all software developed with C-DEBI funding will be open-source and publicly available; it will be maintained on the C-DEBI server, searchable *via* the C-DP and linked from the C-DEBI home page. The C-DP will also provide direct links to appropriate webpages of software authors.

New Data Repository/Archive

The vast majority of C-DEBI data will have a permanent home in the Biological and Chemical Oceanography Data Management Office (BCO-DMO) site (either as a directly deposited data or as a link to another government data repository (i.e., NCBI). Limited C-DEBI resources are best directed first towards getting new data into the repositories, and updating the DP so that search and access are facilitated. However, where no data or sample repository exists for collected data, the DP team will first work with existing data repositories to see if they can make modifications to archive data that currently is not being captured.

If there is no other working alternative, C-DEBI will archive and allow public access to additional data sets. We consider this to be the least desirable alternative because existing public repositories are already structured to maintain data integrity and access for the long term (decades and beyond). That said, we will assure that high-priority C-DEBI data and related products are archived. If we are faced with the dilemma of how to store and make available new data types, we will consult with our community and NSF program officers to find a suitable solution.