NATIONAL HIGH MAGNETIC FIELD LABORATORY

Florida State University
1800 East Paul Dirac Drive
Tallahassee, Florida 32310
nationalmaglab.org

MAGLAB ICR FACILITY USER DATA MANAGEMENT PLAN

Contact: Christopher Hendrickson (hendrick@magnet.fsu.edu)

Funder: National Science Foundation Division of Chemistry and Division of Materials Research

Last modified: 2023-01-10

ABSTRACT:

Ensuring that publicly-funded research data is preserved and freely available safeguards efficient use of government resources and facilitates efficient delivery of scientific discoveries to maximize impact. The National Science Foundation (NSF) supports FAIR (**F**indable, **A**ccessible, **I**nteroperable, and **R**eusable) data guiding principles [1], and considers data management planning as integral to any NSF-funded research. Therefore, products of research generated at MagLab user facilities should be made available to the scientific community and general public. Specifically, this policy requires that all research products generated at the MagLab Ion Cyclotron Resonance (ICR) User Facility be digitally accessible upon publication, or within 3 years. This data management plan (DMP) details resources available to ICR facility users, and outlines procedures for managing data and the products of research in alignment with FAIR principles.

[1] http://www.go-fair.org/fair-principles

PRODUCTS OF THE RESEARCH

Raw data are a single, or a collection of recorded transient signals or mass spectra. Metadata and the products generated by research vary depending on sample and application type. For samples prepared externally, users are responsible for capturing and organizing descriptions of samples, protocols for their preparation, and relevant quantitative and qualitative information about the samples (e.g. experimental conditions, sample quantity and concentration, solvents/buffers, etc). ICR facility personnel will assist users with capture, storage, and organization of ICR experimental conditions, and data and metadata resulting from work taking place at the MagLab. This includes details of any further on-site sample processing prior to data collection, as well as any products of research resulting from data analysis.

DATA FORMAT

Raw data is electronically recorded and stored in the form of *.dat files or Thermo Fisher Scientific *.raw files. The file formats of metadata and products generated by experiments vary depending on raw data and application type. Data, metadata, and the products of research that are stored in unusual or not generally accessible formats will be converted to more accessible formats, or necessary software will be made available to support users and independent investigators with data processing/visualization as described below.

Information provided by users including sample description, preparation protocols, and ICR experiment metadata are recorded and stored in various electronic file formats (e.g. *.raw, *.xlsx, *.docx, *.pptx, *.html, *.txt, or *.pdf) with associated raw data and other products of research.

The facility uses custom-built Predator data stations to collect some ICR data. These data are stored in a custom file format denoted by the *.dat extension. Data collected in the *.dat file format includes the transient signal, which is processed with Predator Analysis software into a mass spectrum. For complex mixture (petroleum, anthropogenic environmental contaminants, and natural organic matter) analysis, each spectrum is calibrated and exported to a peak list (*.pks or *.csv file), with calibration parameters recorded and saved in a *.csv file. Molecular formulae are assigned with PetroOrg© software, and elemental compositions exported and stored as *.xls files.

Predator Analysis, PetroOrg©, and other useful software are available free-of-charge to ICR facility users and data users for noncommercial purposes. Commercial use of PetroOrg requires a license obtained from the Florida State University Office of Commercialization [2]. ICR facility personnel provide training and assistance to new software users upon request. For more information, visit the MagLab's ICR Software webpage [3].

To facilitate data manipulation, data collected in the .raw file format (Thermo Fisher Scientific) are routinely converted to open (vendor-neutral) .mzML files, an XML-based format for mass spectrometer output files that is broadly supported by the mass spectrometry community, industry, and open-source software developers. There are several findable and freely-available .raw file converters [4], and the majority of commercially available software includes embedded converters. Therefore, these data will be made available in their original file format.

For mass spectral imaging (MSI), data are collected as .dat files and phased as part of the data processing workflow. ICR facility staff plan to construct software that converts a collection of *.dat files associated with an image into *.pks files. The combined peak lists will then be converted into a *.mzML file. The *.mzML file in conjunction with an indexing file input from Spectroglyph (MSI sampling/ionization source software) will be converted into an *.imzML file with third party software. The *.imzML file format is the universal MSI file format that is conventionally made publicly available. Raw data is not as useful, and is generally not submitted to imaging repositories. Users will have access to all raw data required to create the *.imzML file, but only *.imzML files are required to be submitted to repositories.

ICR Facility data processing workflows that contain peptide or protein spectral matches (search results) will be made available in their original output file format (e.g. *.tdreport, *.pdResults), and in an appropriate open file format (e.g. *.mzIdentML, tab-delimited *.txt files, *.csv, *.html, etc.) chosen by the PI in consultation with ICR facility staff. These formats will be submitted as Supporting Information, made available by the journal upon publication, and submitted to public repositories along with associated raw data and metadata.

[2] https://www.research.fsu.edu/research-offices/oc/technologies/petroorg-software/
[3] https://nationalmaglab.org/user-facilities/icr/icr-software
[4] http://proteowizard.sourceforge.net/doc_users.html

DATA SHARING AND ACCESS

RESPONSIBILITIES OF THE PRINCIPAL INVESTIGATOR

The Principal Investigator (PI) is the steward of the research data, will select the vehicle(s) for publication or presentation of research products, and will have ultimate authority in their initial use.

Research activities detailed in ICR user proposals and approved for magnet time are expected to result in presentations, publications, or other vehicles for dissemination of data and results. Details of experimental work and metadata (e.g., description of samples, experimental protocols, algorithm specifications, and database schemas) should be included with published data. Published manuscripts should include digital object identifiers (DOIs) and other appropriate persistent identifiers to indicate where relevant data and metadata can be accessed. Users are encouraged to work in collaboration with ICR facility personnel to verify data or results before use in forums such as publications, presentations, and grant or patent applications.

It is the responsibility of the PI to ensure protection of privacy, confidentiality, intellectual property, national security, or other rights or requirements. The PI is encouraged to disclose such requirements to ICR facility staff listed as collaborators to the extent necessary to facilitate compliance. For research involving human subjects, the PI must include the relevant institutional review board (IRB) number(s) in the submitted user proposal, or magnet time cannot be granted. Additionally, the PI must comply with all public access requirements that are laid out by other funding agencies sponsoring the research, in addition to the ICR facility data access policies.

The NSF Public Access Policy requires PIs who publish peer-reviewed journal articles or juried conference papers to make copies of such items (either the final accepted version, or the version of record) available to the public free of charge within one year of publication [5]. The NSF Public Access Repository (NSF-PAR), provides mechanisms that enable NSF-supported investigators to meet this requirement, and provides search mechanisms to enable the public to find and use these materials [6].

[5] https://www.nsf.gov/news/special_reports/public_access/index.jsp
[6] https://www.nsf.gov/news/special_reports/public_access/about_repository.jsp

DATA SHARING PRACTICES

Prior to publication, project data and metadata will be shared with registered MagLab users listed as project collaborators. Requests from other interested parties will be directed to the PI. The PI initial use authority does not control sharing data with ICR facility personnel to gauge instrument performance, meet reporting requirements for the facility, or for preservation and archival purposes.

The MagLab is exploring the Open Science Framework (OSF) to serve as a project management and data sharing platform between ICR facility personnel and external users. Users and staff are encouraged to use the OSF for data transfer, access, and storage, but it is not required. Users can send, receive, and share materials and data on physical media in person, through parcel post, or through their virtual delivery mechanisms of choice in consultation with ICR facility staff.

DATA ACCESS POLICY

This policy applies only to data and metadata collected at the ICR User Facility under the user program. Products of proprietary research, not funded by the NSF, are exempt from these data access requirements. To balance the need to make data openly available to the community with user expectations that they will be able to publish results of their scientific efforts without preemption, data and metadata associated with a user project are expected to be made publicly available when an associated manuscript is published, or within 3 years of the date the project was last assigned magnet time, unless a related publication or patent application is actively under review. Repository entry, DOI, and other relevant accession information should be included in publications and must be reported [7] to the MagLab at the time of publication or conclusion of the data embargo period.

Some data are not required to be made publicly available, because they will not form the basis of publishable research findings nor are associated with a user project. These include data from experiments known to be faulty in some regard, e.g. through mishap or due to a flawed experimental design, data from preliminary experiments that are not intended to be delivered to ICR facility users, standards/calibration runs for which results are not needed to interpret legitimate project data, and data generated to verify successful operation of the instrument or demonstrate capability. Users should consult ICR facility staff regarding the type of data collected and its suitability for public consumption.

[7] https://reporting.magnet.fsu.edu/

PUBLIC DATA REPOSITORIES

FAIR guidelines [1] stipulate that data and associated metadata should be submitted to a discipline-specific, community-recognized, public repository. The project PI is expected to utilize an appropriate repository. Recommended repositories are listed in the table below.

Data-type, Field, or Funding Agency	Repository	Link to homepage
Proteomics	MassIVE (MASS spectrometry Interactive Virtual Environment)	https://massive.ucsd.edu/
Proteomics & Mass Spectral Imaging	PRIDE Archive (PRoteomics IDEntifications Database)	https://www.ebi.ac.uk/pride/archive/
Metabolomics	MetaboLights	https://www.ebi.ac.uk/metabolights/
NSF Division of Ocean Sciences	Biological and Chemical Oceanography Data Management Office	https://www.bco-dmo.org/
NSF Office of Polar Programs	Arctic Data Center	https://arcticdata.io/
Gulf of Mexico Research Initiative	Gulf of Mexico Research Initiative Information and Data Cooperative (GRIIDC)	https://data.gulfresearchinitiative.org/

Marine geochemistry	Marine Geoscience Data System (MGDS)	https://www.marine-geo.org/
Geochemistry	Interdisciplinary Earth Data Alliance (IEDA)	https://www.iedadata.org/
Earth & Environmental Science	PANGAEA	https://pangaea.de/
Natural Products	Global Natural Products Social Molecular Networking (GNPS)	https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp
Earth Sciences	Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE)	https://data.ess-dive.lbl.gov/
Oceanography	Rolling Deck to Repository (R2R)	https://www.rvdata.us/
Terrestrial biogeochemistry	Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC)	https://daac.ornl.gov/
Generalist	Open Science Framework	http://osf.io/

If no suitable, community-recognized resource is available, data and associated metadata should be submitted to a recognized generalist repository. The journal, *Scientific Data* (Springer Nature), recommends several generalist repositories [8]. Among them, the ICR Facility recommends the Open Science Framework (Center for Open Science), a free and open platform for research project management and a reliable data repository [9].

The OSF supports the ability to embargo data and metadata in accordance with the policies outlined above. While embargoed, all submitted materials or datasets are given their own unique, persistent URLs. DOIs can be generated when projects or selected components are made public. These may be cited and accessed by the public, and are indexed in Google Scholar. The OSF is a flexible alternative to some field-specific repositories to efficiently, and wholly disseminate all data and metadata related to complex, large-scale projects spanning multiple disciplines.

Any materials deposited in public data repositories should include the "Policies for Re-use, Re-distribution, and Production of Derivatives" section, below. Data that is submitted to repositories is made available per the terms, conditions, and licenses adopted by the repository.

[8] https://www.nature.com/sdata/policies/repositories#general
[9] https://osf.io/

POLICIES FOR RE-USE, RE-DISTRIBUTION, AND PRODUCTION OF DERIVATIVES

Authors of any publications or presentations that utilize ICR facility data, results, software, or other resources are encouraged to cite relevant literature, include relevant DOIs, or otherwise acknowledge the researchers who generated the samples, data, results, software, or other materials.

In addition, all published manuscripts, datasets, and presentations must acknowledge the MagLab ICR Facility, and facility support (including NSF grant number) as outlined below:

"*A portion of this work was performed at the Ion Cyclotron Resonance User Facility at the National High Magnetic Field Laboratory at Florida State University, which is supported by the National Science Foundation Divisions of Materials Research and Chemistry (DMR-2128556) and by the State of Florida*."

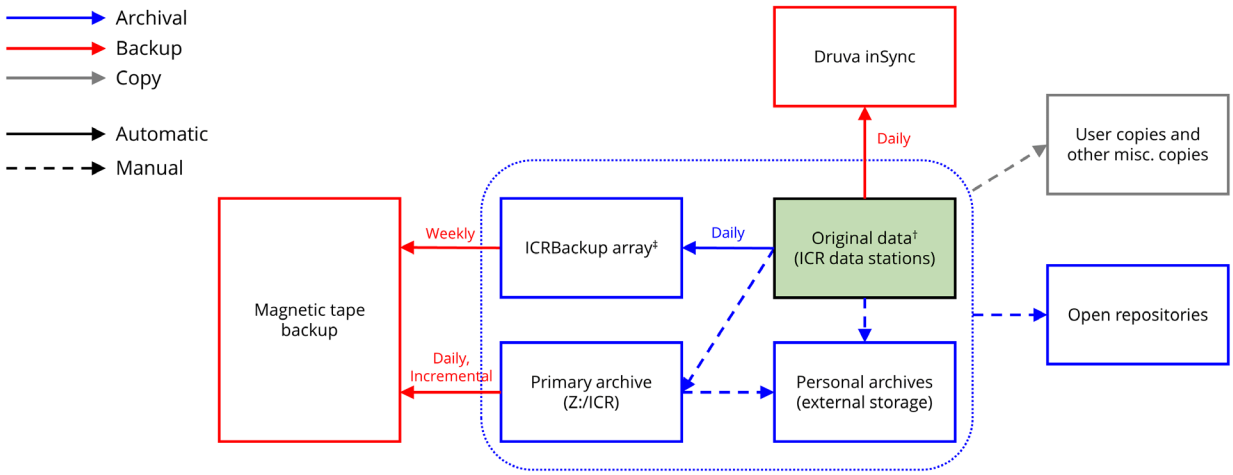
For data collected from 2012-2017, the appropriate grant number is DMR-1157490. For data collected from 2018-2022, the grant number is DMR-1644779. For data collected from 2023-2027, the grant number is DMR-2128556. Please include all grant numbers corresponding to the periods during which data were collected.

DATA ARCHIVE

The ICR facility archives and backs up all raw ICR data. All raw data collection is performed on ICR facility data stations, and data are automatically archived to a dedicated RAID array (ICRBackup) each evening. All raw data can be further archived on additional MagLab storage infrastructure (the Z drive) or using cloud storage (Box, Google Drive, OneDrive, etc.). Data archived on the MagLab Z drive and the ICRBackup array are regularly backed up to magnetic tape to ensure data durability. ICR data stations are also backed up daily using Druva inSync.

All backup servers are managed by the MagLab Computer Support Group. Storage capacity is expected to suffice until at least 2028. All data that has ever been archived has been retained and this will continue indefinitely. PIs are encouraged to utilize any institutional data archival services available to them in addition to MagLab resources.

DATA MANAGEMENT MAP



†: Data stations run Windows 10. Acquired data is stored in C:/Data.
‡: Contents of C:/Data and a few other relevant folders containing instrument methods are automatically archived daily. Read-only access to ICRBackup is available to ICR faculty & staff.