EDRN Informatics and Data Sharing Subcommittee

February 3, 2025

Agenda

1. Previous Action Items
2. NCI EDRN Data Sharing Requirements and Reporting
3. Improving the quality and FAIRness of EDRN data
4. Approach to Address Challenges
    a. Review FAIR Data Guidance for EDRN
    b. Training and Support for Making Data FAIR
5. DICOM Header Standards Working Group Call Update

Action Items

1. JPL to populate the grid of Roles and Responsibilities for FAIR-based data presented on the call to the Public Portal - DONE
2. JPL to promote investigator trainings to ensure that NCI policies are followed - JPL will schedule calls with Collaborative Groups
3. PI's to review their data in LabCAS and let JPL know of any issues.
4. Discuss a roadmap for additional hackathons and workshops.
5. JPL to schedule DICOM Header Standards call with EDRN DICOM Imaging Investigators - DONE. Monthly calls. Next call Thursday, Feb 27 - 12pm PT/3pm ET
6. JPL to review EDRN FAIR Data Guidance Page and Training for each Collaborative Group - reviewing on this call.

Data Sharing and FAIR Data Collection for EDRN - NCI Requirements

Per the EDRN External Data Sharing and Reuse Policy Cycle V, data must be made available for public use.

A primary goal of EDRN data collection is to ensure the reusability of the data by groups beyond those who originally collected it.

Quarterly Reporting

- JPL will provide quarterly updates on EDRN data holdings.
- The 1st Quarterly Report was submitted in December 2024.

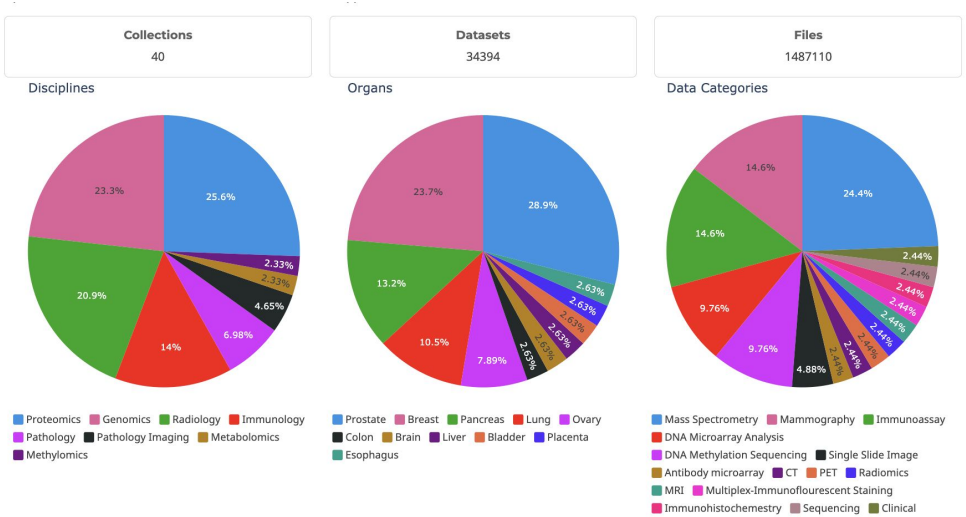
Data Status Reporting will include:

- The status of the data.
- Compliance with the EDRN Data Sharing Policy and Standards.

Non-Compliant Data

- Identified non-compliant data will either be improved to meet standards or returned to the Principal Investigator (PI).

Overview of Data Holdings and Knowledge Environment

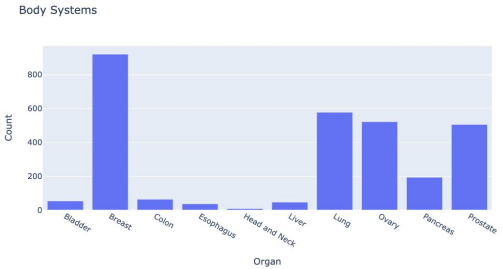


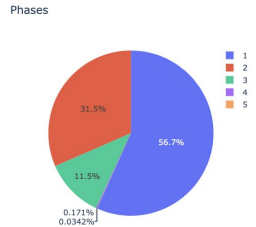
Biomarkers annotated

- 1661 Phase 1
- 923 Phase 2
- 337 Phase 3
- 5 Phase 4
- 1 Phase 5

- 333 protocols
- 2,339 publications

- 40 Data Collections
  - 26 Public Collections (65%)
  - 3 Collections AI ready and alignment with FAIR principals
- 36,628 datasets
- 1.49M files





Key Challenges in Data Submission to JPL

PII Detection

- Uploaded data frequently contains PII, requiring detection, deletion, and re-uploads.

Impact: This process often requires multiple rounds of checks and uploads, requiring significant effort from both sides.

Data Quality Issues

- Empty directories, small set of files and images that appear blank.

Impact: Missing data may go unnoticed until future analysis are done.

Image Header Standards

- DICOM Image headers are not standard between sites.

Impact: This process often requires multiple rounds of checks and uploads, requiring significant effort from both sides.

Incomplete Site/PI or Expert Review

- Sites and/or data experts, often do not review their data after it is published in LabCAS.

Impact: Errors, misrepresentations, and missing data may go unnoticed.

Example: Image Reference Sets

- PII Detection
  - Uploaded images often contain PII, requiring deletion and re-uploads. This process can involve multiple iterations and significant effort from both the site and JPL IC
- Empty Directories and Blank Images in Uploads
  - Images uploaded include directories that are empty or images appear blank.
- DICOM Header Standards
  - Images were uploaded and many sites did not understand how to update the header to change PatientID=EventID.
- Site Review Responsibility
  - Once images are uploaded, it is essential for sites, as the data experts, to review the published data. This ensures the images are usable, correctly organized, and free of errors, misrepresentations, or omissions.

Additional Challenges to Making Data FAIR

Missing Key information that are critical for making data reusable:

- ReadMe File: Explains the data, algorithms, and computations applied to raw data to enable reuse.
- Standard Operating Procedures (SOPs): Descriptions of study procedures for replicability.
- Ancillary Data: Clinical or related data captured during the study.
- Data Dictionaries: Metadata definitions for ancillary data to ensure understanding and FAIR compliance.

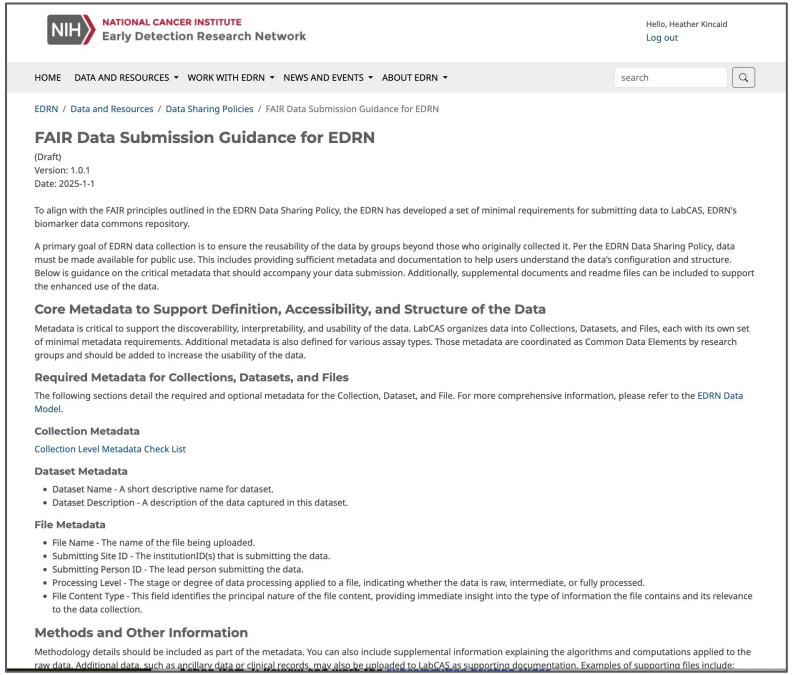
Metadata Gaps

- Metadata is often incomplete, making it challenging to meet FAIR principles (Findable, Accessible, Interoperable, Reusable).

DRAFT - FAIR Data Guidance for EDRN

The FAIR Data Submission Guidance is designed to assist investigators in preparing their data to align with FAIR principles. EDRN Sites must:

- Submit data with core metadata to support its definition, accessibility, and structure.
- Include supporting documentation such as Methods, SOPs, and Data Dictionaries to enhance reusability.
- Follow standard file naming conventions wherever possible.
- Maintain consistency in the organization of files and folders.
- Ensure data usability is verified by the producing site and a scientific investigator.



NIH NATIONAL CANCER INSTITUTE
Early Detection Research Network

Hello, Heather Kincaid
Log out

HOME DATA AND RESOURCES WORK WITH EDRN NEWS AND EVENTS ABOUT EDRN search

EDRN / Data and Resources / Data Sharing Policies / FAIR Data Submission Guidance for EDRN

FAIR Data Submission Guidance for EDRN

(Draft)
Version: 1.0.1
Date: 2025-1-1

To align with the FAIR principles outlined in the EDRN Data Sharing Policy, the EDRN has developed a set of minimal requirements for submitting data to LabCAS, EDRN's biomarker data commons repository.

A primary goal of EDRN data collection is to ensure the reusability of the data by groups beyond those who originally collected it. Per the EDRN Data Sharing Policy, data must be made available for public use. This includes providing sufficient metadata and documentation to help users understand the data's configuration and structure. Below is guidance on the critical metadata that should accompany your data submission. Additionally, supplemental documents and readme files can be included to support the enhanced use of the data.

Core Metadata to Support Definition, Accessibility, and Structure of the Data

Metadata is critical to support the discoverability, interpretability, and usability of the data. LabCAS organizes data into Collections, Datasets, and Files, each with its own set of minimal metadata requirements. Additional metadata is also defined for various assay types. Those metadata are coordinated as Common Data Elements by research groups and should be added to increase the usability of the data.

Required Metadata for Collections, Datasets, and Files

The following sections detail the required and optional metadata for the Collection, Dataset, and File. For more comprehensive information, please refer to the EDRN Data Model.

Collection Metadata

Collection Level Metadata Check List

Dataset Metadata

- Dataset Name - A short descriptive name for dataset.
- Dataset Description - A description of the data captured in this dataset.

File Metadata

- File Name - The name of the file being uploaded.
- Submitting Site ID - The institutionID(s) that is submitting the data.
- Submitting Person ID - The lead person submitting the data.
- Processing Level - The stage or degree of data processing applied to a file, indicating whether the data is raw, intermediate, or fully processed.
- File Content Type - This field identifies the principal nature of the file content, providing immediate insight into the type of information the file contains and its relevance to the data collection.

Methods and Other Information

Methodology details should be included as part of the metadata. You can also include supplemental information explaining the algorithms and computations applied to the raw data. Additional data, such as ancillary data or clinical records, may also be uploaded to LabCAS as supporting documentation. Examples of supporting files include:

FAIR Data Submission Guidance

Roles and Responsibilities for making Data FAIR

Responsibility	Investigators	JPL	NCI	DMCC	Future Needs
(F)indable	Submit data to JPL in a timely manner along with metadata; verify completeness	Capture, ingest, and post data and metadata so it is accessible; catalog metadata for search	Enforce that data is delivered per the data management plan	Data tracking for EDRN Validation and Reference Set studies	Require **all data is delivered per a data management plan** (completeness)
(A)ccessible	Identify sensitivity and sharing of data (e.g., public vs study)	Configure access to data based on sensitivity of data; ensure it can be accessed by the right groups	Enforce that data is made public per the NCI policy.	Notify JPL of Validation and Reference Set data and metadata.	Require that **data be made public** per NCI policies
(I)nteroperable	Structure data so it can be integrated with other data; Validate data when posted to LabCAS	Provide API and tool access for data integration; support search; Link with other data via metadata	Enforce that meets specific structure requirements, where applicable	Support statistical analysis of validation study data, including structure of data	Enforce that **data standards are followed** including appropriate metadata and structured data
(R)eusable	Structure data for reuse Validate data in LabCAS Need to deliver any required software Provide documentation	Provide ability to run central pipelines to support repeatable processing of derived data	Enforce sites meet usability process/requirements	Support analysis of validation study data	Include a process to ensure **usability of data** and metadata (e.g., peer review to validate quality, a reusability test); capture documentation and software

Proposed Solutions to Help Prevent PII in Initial Uploads

Provide De-identification Resources:
Add the Health and Human Services **- Methods for De-identification of PHI** website as a reference in our portal and processes.

Enhance Metadata Requirements:
Include a De-identification Method Expert Determination and Safe Harbor[1] in collection metadata with these two options:

1. Expert Determination:
   A qualified expert certifies that the risk of re-identification is very small, documents the methods used, and justifies the determination.
2. Safe Harbor:
   Specific identifiers (e.g., names, addresses) are removed as per the Privacy Rule (see full list on HHS website).

1. The Health Information Technology for Economic and Clinical Health (HITECH) Act was enacted as part of the American Recovery and Reinvestment Act of 2009 (ARRA). Section 13424(c) of the HITECH Act requires the Secretary of HHS to issue guidance on how best to implement the requirements for the de-identification of health information contained in the Privacy Rule.

Key “New Data Submission” Steps for EDRN Sites

- ❑ Understand the FAIR data submission requirements:
  - ❑ Review EDRN FAIR Guidelines
  - ❑ Review Existing FAIR and “Data Ready” Collection example in LabCAS
- ❑ Obtain required metadata
  - ❑ Use Checklist for Collection Level Metadata
- ❑ Review metadata and data annotations with JPL Informatics Center (JPL IC)
- ❑ Prepare FAIR-aligned Collection Level metadata and submit
  - ❑ Use Collection Metadata Submission Form

Uploading data to LabCAS -> Continue with steps below:

- ❑ Prepare your data files- Review the following sections of the FAIR Data Submission Guidance for EDRN:
  - ❑ De-identification of Data - ensure all data and images are fully de-identified before upload.
  - ❑ Data to Upload - identify the necessary data.
  - ❑ Organization of Files and Folders - organize the data into a logical folder structure
- ❑ Review data organization with JPL IC
- ❑ Upload Data (Request Aspera Account if necessary)
- ❑ Work with JPL IC to determine if expert review is necessary
- ❑ Review & Verify data published in LabCAS

See New Data Submission Procedure for detailed steps for all stakeholders

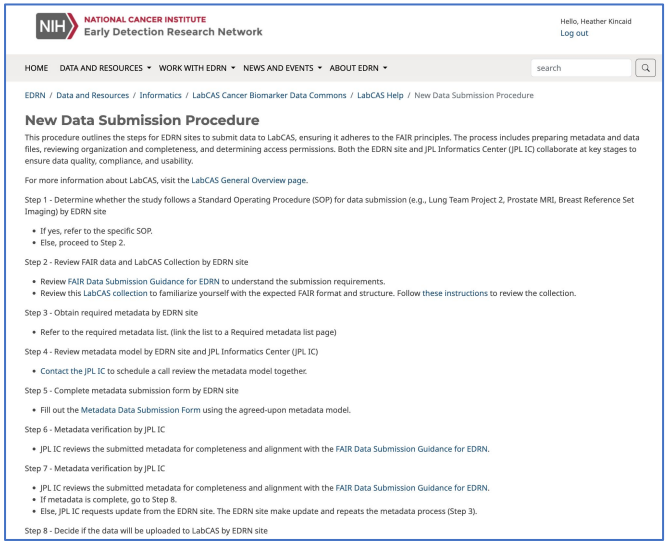
Key “Update Existing Data” Steps for EDRN Sites

- ❑ Understand FAIR data submission requirements
  - ❑ Review EDRN FAIR Guidelines
  - ❑ Review Existing FAIR and “Data Ready” Collection
- ❑ Identify gaps in metadata based on Guidelines
- ❑ Obtain required metadata
  - ❑ Use Checklist for Collection Level Metadata
- ❑ Review metadata and data annotations with JPL Informatics Center (JPL IC) if there are any metadata gaps
- ❑ Prepare FAIR-aligned collection metadata and submit metadata
  - ❑ Use Collection Metadata Submission Form
- ❑ Review data and organization with JPL IC if there are any data gaps
- ❑ Prepare your data - Review the following sections of the FAIR Data Submission Guidance for EDRN:
  - ❑ **De-identification of Data - ensure all data and images are fully de-identified before upload.**
  - ❑ Data to Upload - identify the necessary data.
  - ❑ Organization of Files and Folders - organize the data into a logical folder structure
- ❑ Review data organization with JPL IC
- ❑ Upload Data (Request Aspera Account if necessary)
- ❑ Work with JPL IC to determine if expert review is necessary
- ❑ Review & Verify data published in LabCAS

See Update Existing Data Submission Procedure for detailed steps for all stakeholders

Training and Support Plans

- Contact each of the EDRN Collaborative Groups to provide training on an upcoming monthly call
  - Specific training materials to walk through using documentation and guidance posted on EDRN public portal:
    - FAIR Data Submission Guidance
    - Submitting New Data Steps
    - Updating Existing Data Steps
  - Share end to end data submission workflows
- Provide documentation, Guidance and Recorded Training on EDRN Public Portal
- Provide additional training call and support as needed



NIH NATIONAL CANCER INSTITUTE
Early Detection Research Network

Hello, Heather Kincaid
Log out

HOME | DATA AND RESOURCES | WORK WITH EDRN | NEWS AND EVENTS | ABOUT EDRN | search

EDRN / Data and Resources / Informatics / LabCAS Cancer Biomarker Data Commons / LabCAS Help / New Data Submission Procedure

New Data Submission Procedure

This procedure outlines the steps for EDRN sites to submit data to LabCAS, ensuring it adheres to the FAIR principles. The process includes preparing metadata and data files, reviewing organization and completeness, and determining access permissions. Both the EDRN site and JPL Informatics Center (JPL IC) collaborate at key stages to ensure data quality, compliance, and usability.

For more information about LabCAS, visit the LabCAS General Overview page.

Step 1 - Determine whether the study follows a Standard Operating Procedure (SOP) for data submission (e.g., Lung Team Project 2, Prostate MRI, Breast Reference Set Imaging) by EDRN site

- If yes, refer to the specific SOP.
- Else, proceed to Step 2.

Step 2 - Review FAIR data and LabCAS Collection by EDRN site

- Review FAIR Data Submission Guidance for EDRN to understand the submission requirements.
- Review this LabCAS collection to familiarize yourself with the expected FAIR format and structure. Follow these instructions to review the collection.

Step 3 - Obtain required metadata by EDRN site

- Refer to the required metadata list. (link the list to a Required metadata list page)

Step 4 - Review metadata model by EDRN site and JPL Informatics Center (JPL IC)

- Contact the JPL IC to schedule a call review the metadata model together.

Step 5 - Complete metadata submission form by EDRN site

- Fill out the Metadata Data Submission Form using the agreed-upon metadata model.

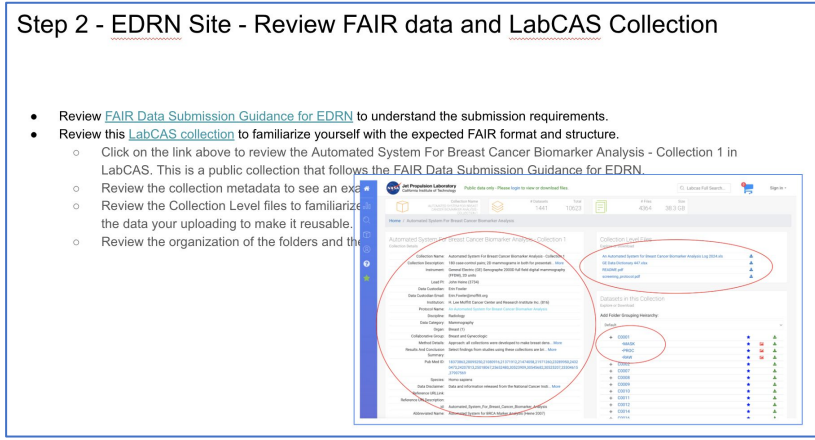
Step 6 - Metadata verification by JPL IC

- JPL IC reviews the submitted metadata for completeness and alignment with the FAIR Data Submission Guidance for EDRN.

Step 7 - Metadata verification by JPL IC

- JPL IC reviews the submitted metadata for completeness and alignment with the FAIR Data Submission Guidance for EDRN.
- If metadata is complete, go to Step 8.
- Else, JPL IC requests update from the EDRN site. The EDRN site make update and repeats the metadata process (Step 3).

Step 8 - Decide if the data will be uploaded to LabCAS by EDRN site



Step 2 - EDRN Site - Review FAIR data and LabCAS Collection

- Review FAIR Data Submission Guidance for EDRN to understand the submission requirements.
- Review this LabCAS collection to familiarize yourself with the expected FAIR format and structure.
  - Click on the link above to review the Automated System For Breast Cancer Biomarker Analysis - Collection 1 in LabCAS. This is a public collection that follows the FAIR Data Submission Guidance for EDRN.
  - Review the collection metadata to see an exa
  - Review the Collection Level files to familiarize
    the data your uploading to make it reusable.
  - Review the organization of the folders and the

DICOM Header Standards Working Group Call Update

First call held January 23, 2025 - great turnout! Approx 20 participants (Prostate, Lung, Breast and Liver)

- Discussed:
  - Drafting a set of minimum required DICOM header tags applicable to all DICOM images.
  - Define modality-specific required header DICOM tags
- Purpose:
  - Support standardized imaging SOPs.
  - Ensure images comply with FAIR principles
- Next Steps
  - Review current EDRN DICOM tags and map them to the TCIA de-identification tags.
  - Radka at University of Miami, Chad He, Yoga from Moffitt and Ashish will meet this week to discuss and prepare a proposal to develop a Macro to standardize PMRI images, as there may be some additional funding for this activity.

Future Calls will be held monthly (4th Thursday of the month at 12pm PT/3pm ET)

  - Next call - Thursday, Feb 27 at 12pm PT/3pm ET

Summary

- Update EDRN FAIR Guidance based on Data Sharing and Informatics Subcommittee feedback
- Update De-identification process based on subcommittees feedback
- Publish all updated FAIR Guidance, updated data submission and updated LabCAS documentation to EDRN Portal
- Schedule training calls with all Collaborative Groups
- Submit Posters for the EDRN Scientific Workshop in March 2025
  - From JPL IC:
    - EDRN FAIR Guidance and LabCAS Data Submissions
    - EDRN Data Resources Available (EDRN Knowledge Environment)
    - LabCAS
    - AI/ML Workshop/Hackathon Progress Updates
- DICOM Standards Working Group Established