

# EDRN Informatics and Data Sharing Subcommittee

February 13, 2023

# Agenda

- Review data sharing policy
- Review subcommittee responsibilities
  - The objectives of the Data Sharing and Informatics Subcommittee are to establish guidelines for the core EDRN data model, core common metadata and to provide a forum for bioinformatics/data scientists within EDRN to collaborate on research pertinent to EDRN.
  - Responsibilities
    - Update the EDRN Data Sharing Policy
    - EDRN Data Sharing Framework – define what data and metadata should be captured in EDRN Cancer Biomarker Data Commons
    - Collaborate with EDRN sites to capture data, algorithms, etc. so that scientific findings can be clearly documented and reproduced
    - Host ML pipelines to allow other sites to use
    - Collaborate with the EDRN Statistical Working Group
    - Develop bioinformatic resources for EDRN members, such as what tools to use for de-identification of DICOM images
    - Develop methods to promote data sharing with NCI programs such as CGAP, CGN, SEER, etc.
    - Support reuse and sharing of bioinformatic tools and capabilities
  - Upcoming EDRN Steering Committee Meetings and agenda
- EDRN data standards/common data elements status and needs
- Data repository support and guidance
- Find inventory of published data from past cycle EDRN data
- Schedule presentations every two months from each funded site by investigators, collaborators, or invited speakers
- Co-chair volunteer

# EDRN Data Policy

Please provide final comments in the proposed EDRN data policy

<https://docs.google.com/document/d/1fEewRaHdf-rqky2qy6zgQ3NeA7hAnrgH/edit#>

# Responsibilities of the Subcommittee

The objectives of the Data Sharing and Informatics Subcommittee are to establish guidelines for the core EDRN data model, core common metadata and to provide a forum for bioinformatics/data scientists within EDRN to collaborate on research pertinent to EDRN.

Responsibilities:

- Update the EDRN Data Sharing Policy
- EDRN Data Sharing Framework – define what data and metadata should be captured in EDRN Cancer Biomarker Data Commons
- Collaborate with EDRN sites to capture data, algorithms, etc. so that scientific findings can be clearly documented and reproduced
- Host ML pipelines to allow other sites to use
- Collaborate with the EDRN Statistical Working Group
- Develop bioinformatic resources for EDRN members, such as what tools to use for de-identification of DICOM images
- Develop methods to promote data sharing with NCI programs such as CGAP, CGN, SEER, etc.
- Support reuse and sharing of bioinformatic tools and capabilities

Upcoming EDRN Steering Committee Meetings and agenda

# Data Standards

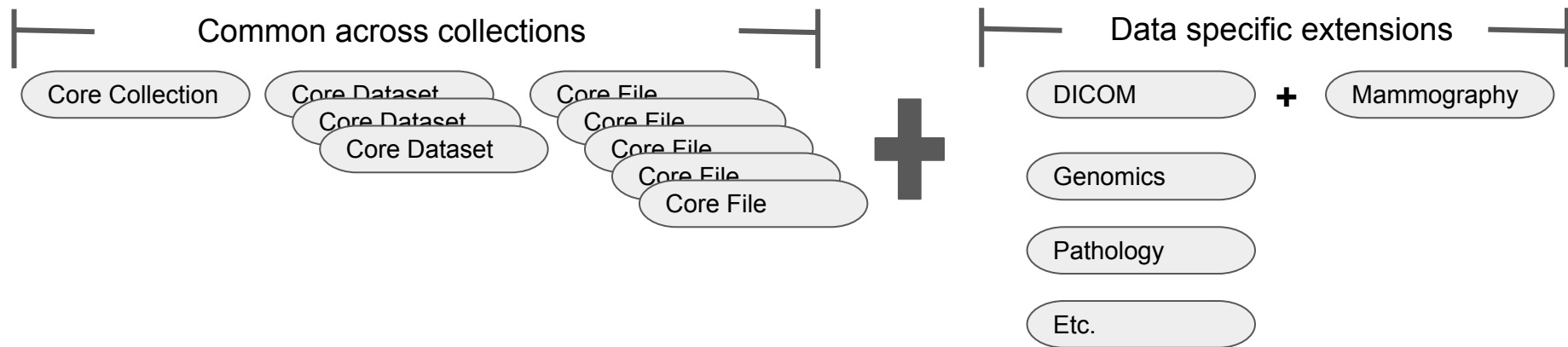
- Common Data Elements
- Metadata standards for loading data into LabCAS
- Discussion: What additional guidance would be useful?

See: <https://edrn.nci.nih.gov/data-and-resources/cde/>

# Data Repository Support and Guidance

- JPL provides “LabCAS” as a Biomarker Data Commons
- LabCAS provides
  - Multi-level security for data sharing within EDRN
  - Plug-in “data pipelines” for processing data on ingestion
  - Capture of metadata and data
  - Organization of data into collections, datasets, and files
  - Image viewing capabilities
  - APIs for downloading data
- JPL can support your site to capture your datasets

# Structuring data for loading into LabCAS



# EDRN Data Upload Process into LabCAS (DRAFT)

## General Upload Process:

1. Collaborate with JPL to determine Core Collection and Dataset(s) hierarchy and metadata
  - a. Contact the Informatics Center at [ic-data@jpl.nasa.gov](mailto:ic-data@jpl.nasa.gov) to upload new data collections
2. Determine data security classification - Public, Private (who can have access)
3. Review [core and discipline](#) specific EDRN metadata dictionaries and templates
4. Create any additional specific metadata extensions
5. Download metadata templates
6. Upload paired data files and metadata
7. IC review and QC uploaded files
8. Data ingest pipeline run to ingest into LabCAS
9. Data QC'd in LabCAS by submitter

*Need to determine when a DOI is assigned*

We plan to move this onto the EDRN Public Portal, when ready



# Inventory of datasets

- NCI is interested in ensuring we can “track” datasets whether delivered to LabCAS or to another repository (e.g., dbGAP)
- EDRN is able to track and catalog datasets delivered to other repositories and provide the linkage to the data
- Discussion: how do we track (1) new datasets generated at EDRN funded labs; and (2) identify existing datasets that have been delivered to other repositories?

# Subcommittee Presentations

We would like to schedule presentations to this group on bioinformatics and data science capabilities at sites every two months

Please send suggestions to [Heather.Kincaid@jpl.nasa.gov](mailto:Heather.Kincaid@jpl.nasa.gov)

# Co-Chair

We are still open to a co-chair to help support the subcommittee.

Please send names to [Dan.Crichton@jpl.nasa.gov](mailto:Dan.Crichton@jpl.nasa.gov)