Breakout Session 4: Track B

Leveraging Intramural NCI Data Platforms for Accelerated Data Sharing

Dr. Janelle Cortner (Moderator)
Director, Data Management and Analysis Program, CBIIT, NCI

Leveraging Intramural Data Platforms for Accelerated Data Sharing

Janelle Cortner, PhD

Data Management and Analysis Program, NCI, CBIIT, OCIO

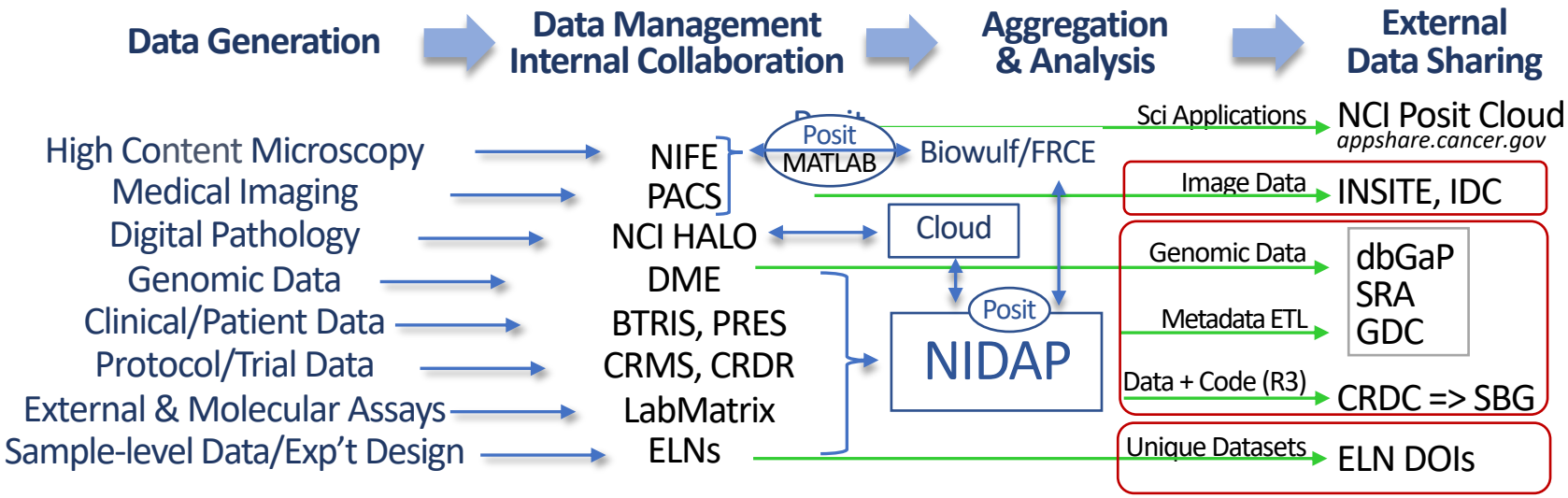
NIH NATIONAL CANCER INSTITUTE

January 17, 2024

DATA MANAGEMENT AND ANALYSIS PROGRAM, NCI, CBIIT, OCIO

Janelle Cortner, PhD

DMAP develops digital research infrastructure aimed at accelerating multimodal data management and analysis across the data life cycle



- IRP data platforms enable Investigators to manage studies with multiple data types & access computational resources
- IRP data platforms can be leveraged to lower the barriers for data sharing (collaboration & secondary reuse)

Leveraging IRP Data Platforms for Accelerated Data Sharing

FY22 High Value Dataset Awards catalyzed development data sharing in 3 major areas:

Leveraging ELNs:

- Efficient Sample-level Metadata Capture
- External Sharing of Unique Datasets via Digital Object Identifiers (DOIs)

Streamlined Imaging Data Sharing via Platform Interoperability

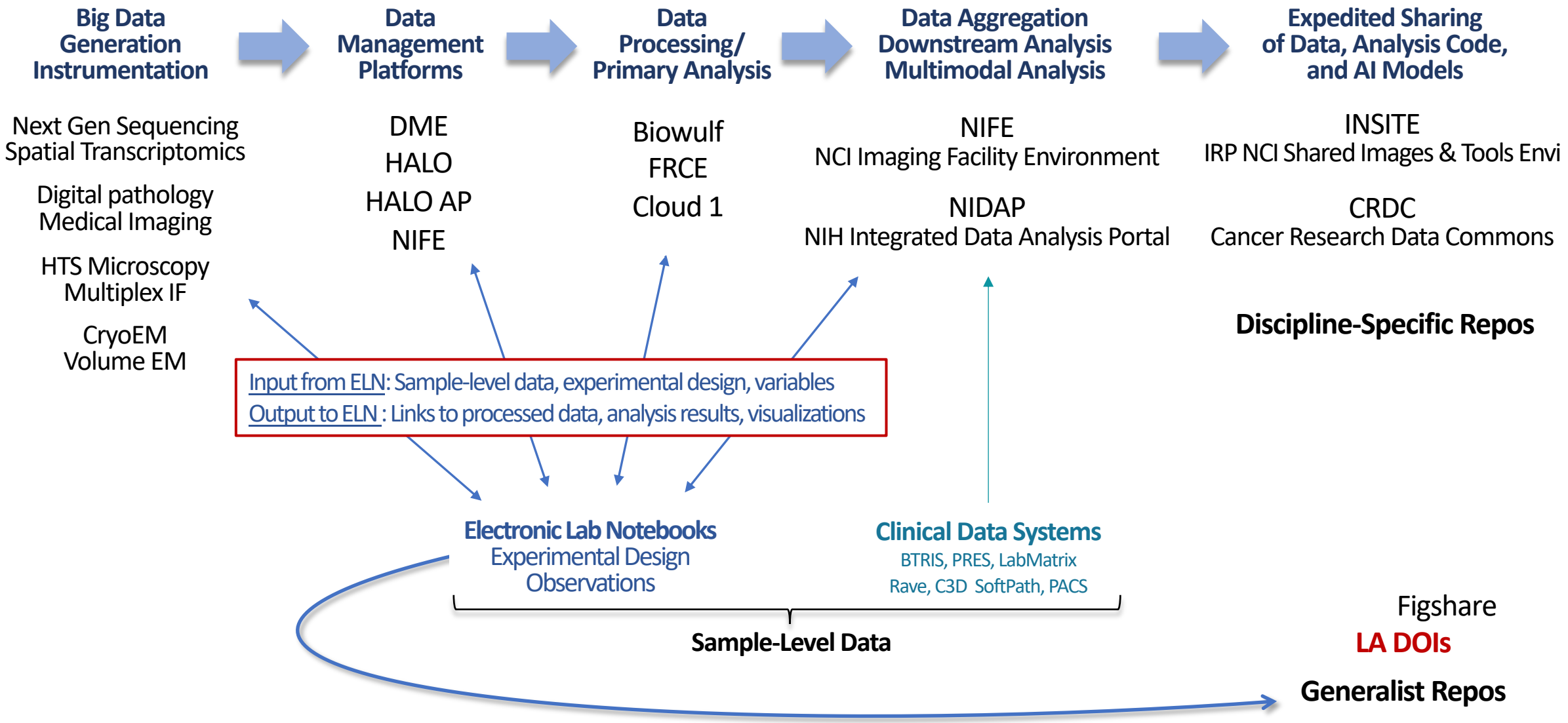
- Uniform metadata hierarchies
- INSITE: Intramural NCI Shared Images and Tools Environment

Accelerating Sharing of Genomic Data:

- Sample-level metadata transformation pipelines in NIDAP
- Platform-mediated Data Transfer from DME to major genomic repos
- Reproducible Research Repository (R^3): Data + Code hosted in a runtime environment

ELNs Serve as the Documentation Hub within an Integrated Data Environment

ELNs facilitate uniformity of sample-level metadata and persistent association with cognate instrument data



Leveraging the LabArchives ELN to Accelerate Sharing of Unique Datasets

Problem: Many journals require that raw data be publicly accessible => reproducibility & reuse
Highly reusable data have discipline-specific repos … but many datasets lack a suitable repository

ELN Solution: Share data directly from the Investigator's LabArchives account in cases when the data:

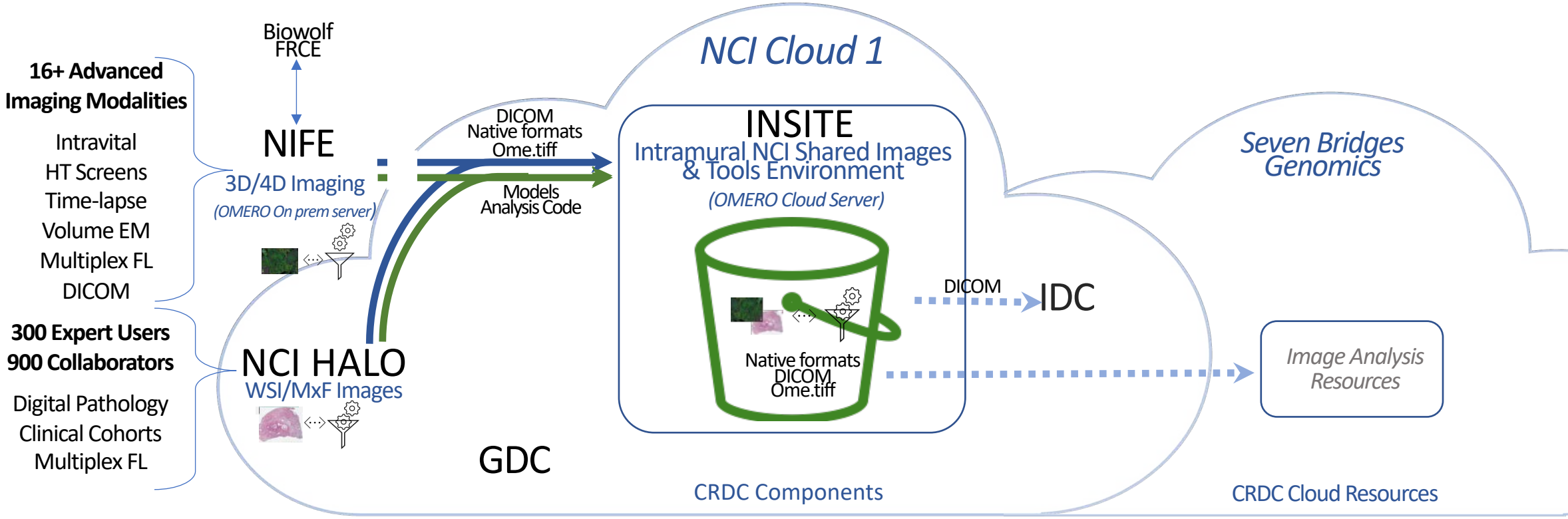
- Do not have a good fit with a discipline-specific repository
- Can be made publicly available without access restrictions

Approaches:

LA Figshare Integration	**LA Digital Object Identifiers**
Initiated directly from the ELN	Initiated directly from the ELN
Creates persistent & searchable DOIs	Creates persistent & searchable DOIs
Creative Commons Licenses: CC0 & CC BY	Creative Commons Licenses: CC0 & CC BY
No paywall; Free downloads w/o logging in	No paywall; Free downloads w/o logging in
20 GB max per uploaded file	1 TB max per shared DOI; no data transfer
20 GB storage for free account + Paid Tiers: 100 GB $ 395 1 TB $ 2,500 5 TB $11,860	Unlimited storage for Enterprise accounts

NCI IRP Imaging Platforms

Leveraging Intramural Data Platforms to Streamline Data Submission and Accelerate Data Sharing



NCI Imaging Facilities Environment (NIFE) Metadata

OMERO db

Metadata intake GUI

Study

- Study Type
- Study Description
- Principle Investigator
- Dataset info (publications etc)

1:n

Study Component

- Imaging Facility (dropdown)
- Principal Investigator ID
- Project Scientist
- Dataset unique ID
- Imaging method (facility-specific dropdown)
- Study/Dataset description
- Supporting documentation links

1:n

Biosample

- Biosample ID
- Biosample description
- Organism (dropdown)
  - GEM strain/ID_____ OR
  - Tissue or Cell Line
    - Cell line _____ OR
    - Tissue/site (dropdown)
      - Organoid or PDX
      - Subsite (dropdown)
      - Normal or Disease/Condition
        - Disease/Condition_____
      - If neoplasm=> Classification *(IDC-O-3, OncoTree)*
  - Intrinsic variables – e.g. genotype
  - Extrinsic variables – e.g. treatments
  - Experimental variables

1:n

Specimen

- Specimen ID
- Location method (e.g. plate/well, block/slide)
- Experimental Status: Test or control
- Sample preparation method
- Other Prep details (e.g. photobleaching) free text
- Channel content: # channels

One entry per channel for:

- Signal/contrast mechanism
- Stain
- Molecular entity detected

1:n

<Image files>

- Image ID
- Image type (e.g. primary, processed, segmented)
- Format and compression
- Magnification
- Dimension extents (pixels^3; tilts)
- Image size description
- Pixel/voxel size description
- Channel representation
- Image processing methods
- QC details, score

- Server path
- Instrument attributes
- Image acquisition parameters

Legend:

Black: Required fields provided by the requestor

Blue: Optional fields

Green: Provided by the Facility or derived/added later for multi-part studies

Genomics Data Sharing: Metadata Prep & Platform Mediated Data Transfer

Leveraging the NIDAP-DME Integration and Unique Platform Capabilities

Sample-level Metadata Preparation

- NCI genomic data are stored in DME along with sample-level metadata in DME's iRODS db
- Additional sample-level metadata from multiple sources integrated with NIDAP are pulled in
- Metadata are mapped to repository templates using NIDAP ETL capabilities in Pipeline Builder
- Transformed metadata are submitted by the user following QC

Platform-mediated Data Transfer of Genomic Data Files from DME to Repositories

- DME => dbGaP: Genomic data files are transferred to the dbGaP Aspera endpoint
- DME => SRA: Genomic data files are pushed to a DME S3 bucket, and then pulled in by SRA

Genomics Data Sharing: Reproducible Research Repository

Date + Analysis Code Shared in an NCI-hosted Runtime Environment in Seven Bridges Genomics

Problem: Currently, data and analysis code are deposited into separate repositories to meet sharing requirements

- Sharing is burdensome for the submitter
- Reproduction is challenging; data & code are downloaded separately then re-deployed in user's infrastructure

Solution: Host NIDAP workbooks with versioned data & code at CRDC => **Reproducible Research Repositories (R3s)**

- The burden of data sharing and submission is reduced
- Reproducibility barriers are eliminated
- Exact version of data and code are preserved (R3)
- A dynamic version in which parameters can be adjusted and user data added (R3D)