CCDH Quarterly Report

Q1 2021

CENTER *for*
CANCER DATA
HARMONIZATION

ccdh.cancer.gov

CCDH Quarterly Report to Federal Oversight (NIH/NCI and FNL)
Date: April 14, 2021

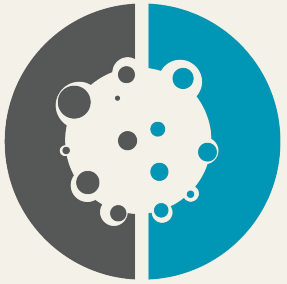
These slides: bit.ly/ccdh-q1-2021

Outline

- Developments for our Community
- Updates to the CRDC-H harmonized data model
- Terminological alignment & services for CRDC nodes
- Tools for putting CRDC-H into use
- Timeline Update & Summary of Planned Phase II Deliverables

Developments for our Community

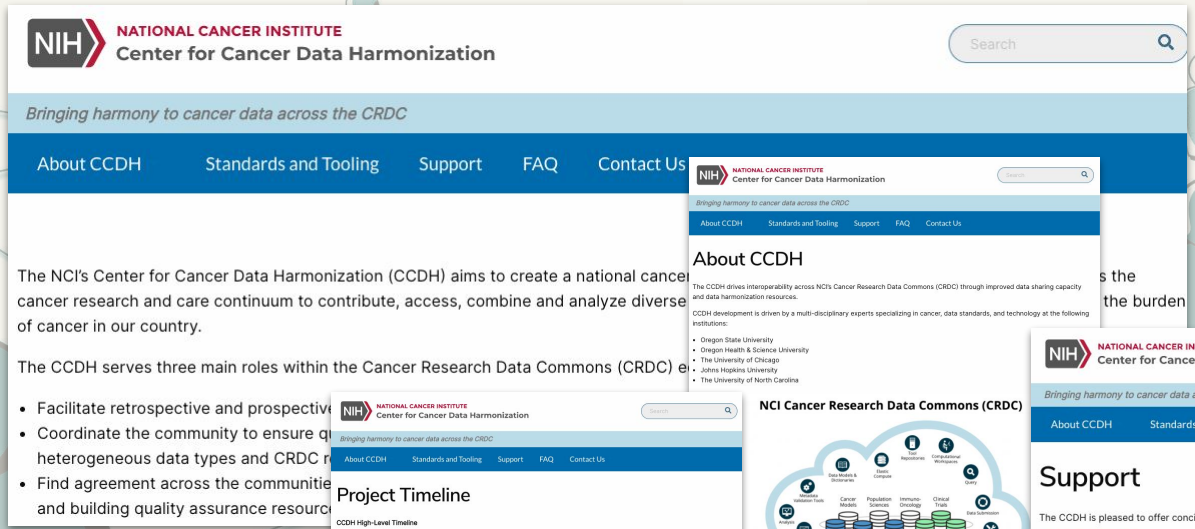
(Nicole & Sam)

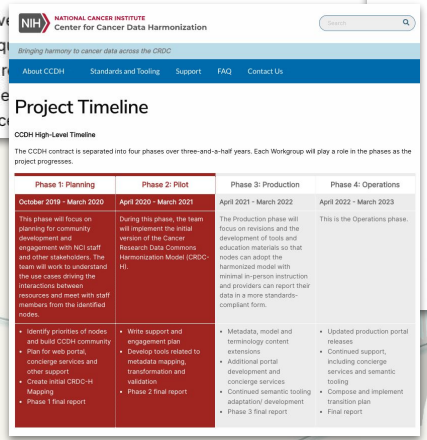


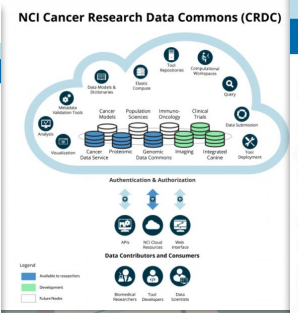
CENTER *for*
CANCER DATA
HARMONIZATION

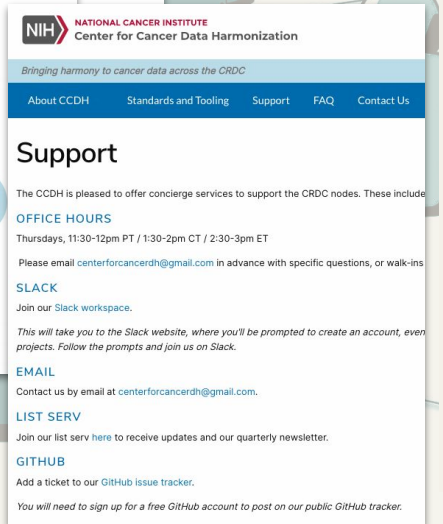
CCDH Website maintenance

ccdh.cancer.gov









- Updates based on community feedback
- Next up: collection of links to CCDH visualization tools & documentation for users

Concierge Services

- Complete: End-to-End Analysis Report
- Ongoing: User Support and Engagement Plan
- Ongoing: Working with technical workstreams to develop documentation
- Maintenance: Help Desk
- Maintenance: *Homing in on Harmonization* sessions
- Maintenance: Quarterly CCDH Newsletter:
  - Includes updates on progress towards deliverables, summaries of presentations, introductions to the CCDH team, and any additional relevant announcements or calls for feedback
- Maintenance: Engagement with nodes, CRs:
  - Meetings with CDS, CTDC, and FHIR Research Gap Analysis

User Support and Engagement Plan

- Overview
- Help desk
- Outreach
  - Newsletter
  - Community interactions
  - GitHub issue tracker
  - Services to CRDC
  - CDA and CCDH Collaboration
- Web Portal Information
- Presentations and Reporting
- Visualizing Harmonization (next slide)

Status: in progress

Draft document: bit.ly/supportPlan

Visualizing Harmonization

- Helping users understand what harmonization looks like:
  - Visualizations:
    - CRDC-H model browser
    - Integrated NCIt-Plus terminology navigator
  - Documentation and tutorials:
    - Transforming Data to CRDC-H
    - Validating data using the CCDH validation tool

Garner requirements and feedback on
the documentation and visualization tools

Homing in on Harmonization Sessions

Summary:

The *Homing in on Harmonization* sessions were developed as a replacement for the bi-weekly office hours to create discussion around a specific topic and increase attendance.

Goal:

To encourage engagement within the CCDH community by holding quarterly sessions incorporating a relevant presentation and an open discussion.

- First session will be held on **Thursday, April 29 at 11 am PT / 1 pm CT**
- Topics are being solicited from the internal team via google form
- Speakers and discussion topics will be determined after suggested topics are received

End-to-end Requirements Analysis

Report: bit.ly/crdc-e2e

Goals

- Understand information flow for each node, including opportunities
- Answer the questions:
  - What standards are being used?
  - How do we structure the resources?
  - Where are the data being shared?

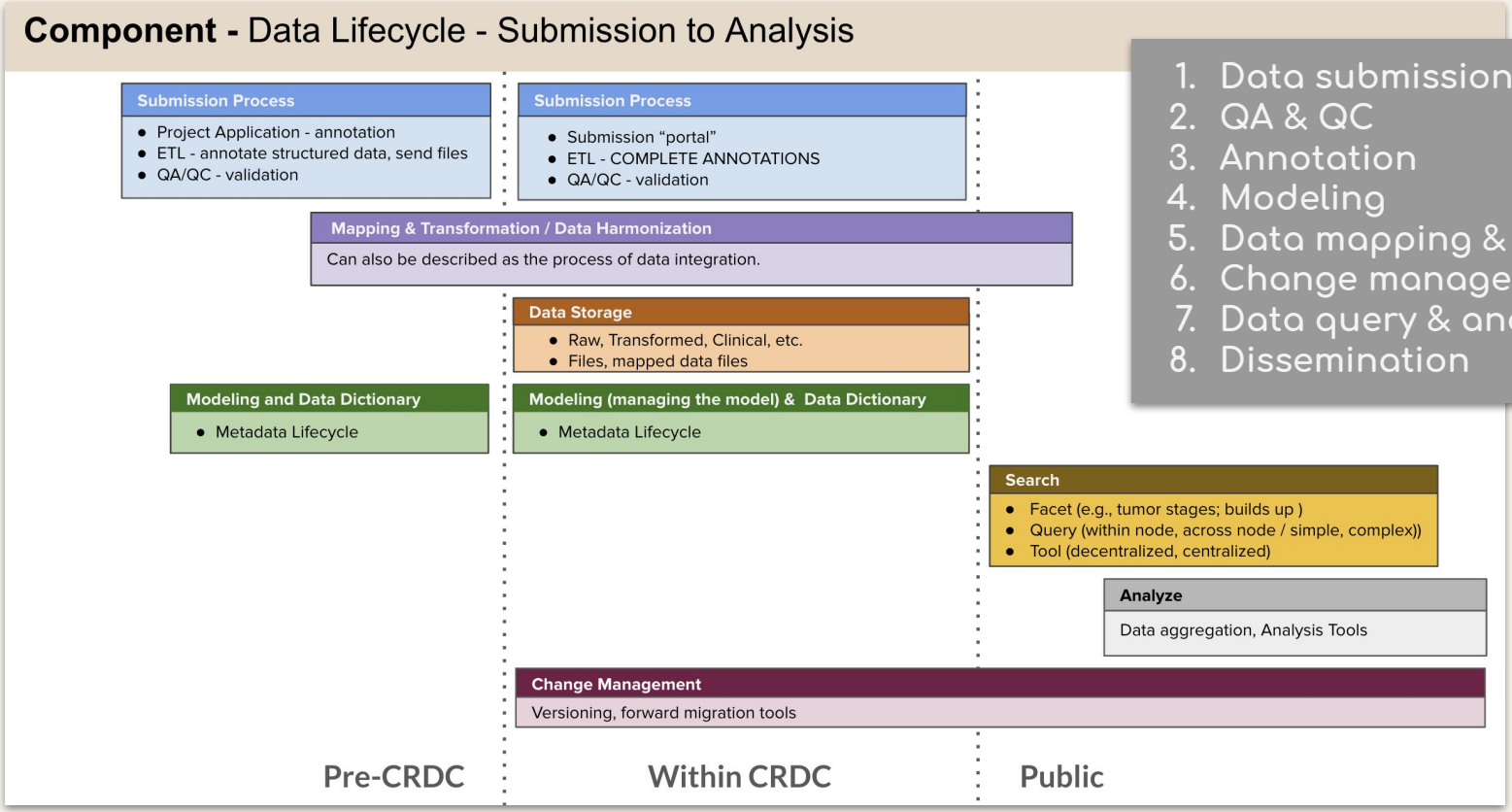
- Prioritize CCDH development based on the greatest needs and best ROI
- Issue formal recommendations to NCI to inform data life cycle management, information architecture, and strategies to better harmonize CRDC resources for improved research analytics.

Stakeholders

NCI leadership, CRDC leadership, nodes and data resource stakeholders, CCDH and CDA collaboration, Other NCI programs.

End-to-end Requirements Analysis

Report: bit.ly/crdc-e2e

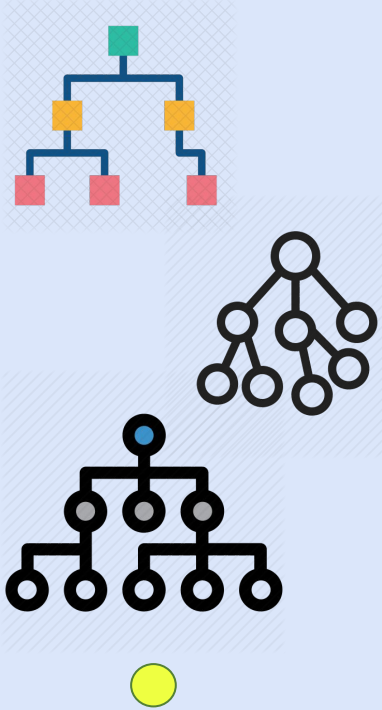


1. Data submission to the nodes
2. QA & QC
3. Annotation
4. Modeling
5. Data mapping & transformation
6. Change management
7. Data query & analysis
8. Dissemination

CRDC data mapping is inconsistent, absent, and/or lossy

Report: bit.ly/crdc-e2e

Source terminologies



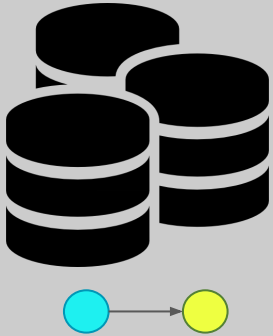
"Mappers"

NIH NATIONAL CANCER INSTITUTE Enterprise Vocabulary Services

ISO caDSR

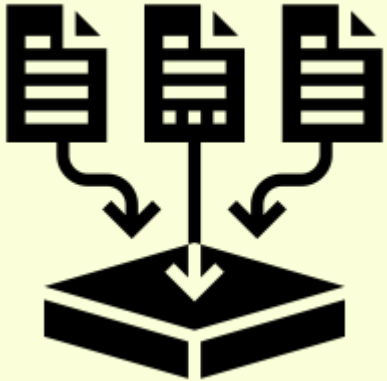
UMLS Unified Medical Language System

Coded Data



Unencoded/locally coded data

Codesets/valuesets to unify data



{ }

How does the E2E inform CCDH work?

Report: bit.ly/crdc-e2e

Data submission

CCDH can improve CRDC interoperability by assisting nodes with data ingestion:

- Terminology services for term lookup, codeset validation, & data transformation
- Create CRDC-H based metadata templates (e.g. using CEDAR) that use CCDH terminology services; these can be extended for node-specific needs

QA & QC

CCDH can develop a suite of services for compliance and QA checking:

- Can be locally deployed or called as services
- Especially relevant if nodes want to deploy CRDC-H compliant APIs
- Check for value set validation, identifier equivalency, synonyms, etc.

How does the E2E inform CCDH work?

Report: bit.ly/crdc-e2e

Annotating Data & Metadata

CCDH can help with:

- Facilitating x-CRDC development of metadata & common identifier standards.
- Facilitate Privacy-Preserving Record Linkage (PPRL) strategies (e.g. for multimodal analytics, patient deduplication, etc.) based upon the foundation laid out with NCI TCIA and NCATS

Mapping and Harmonization

CCDH can help harmonize data across nodes:

- Provide concierge services to help nodes create CRDC-H and CCDH terminology compliant APIs even if they are not natively ingesting data using compliant formats
- Create harmonization strategies and tools support full mapping of fields and value sets to CRDC-H and CCDH terminologies

How does the E2E inform CCDH work?

Report: bit.ly/crdc-e2e

Data Dictionary & Data Modeling

- Provide feedback to nodes on their documentation that will ultimately help their ingests be more interoperable
- Publish CRDC-H model in LinkML; build and improve tooling, particularly in the areas of model and data validation and interoperability with existing data dictionaries, such as the Cancer Data Standards Registry and Repository (caDSR).

Search & Analysis

- Provide recommendations to nodes regarding identifier strategies
- Provide search technologies, semantic best practices, and terminology services to CDA that can support semantic search across data sources encoded or mapped to CRDC-H

How does the E2E inform CCDH work?

Report: bit.ly/crdc-e2e

Cloud Resources

- Work closely with cloud resources following May CRDC-H release to have them evaluate the CRDC-H and provide granular and iterative feedback

CRDC Change Management

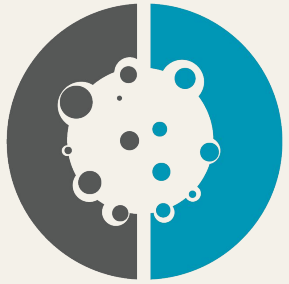
- CCDH helpdesk and office hours will continue to be available for any data model, terminology, or tool requests or questions
- Discuss, document, and communicate change to CRDC resources

CCDH Change Management

- CCDH will provide documentation for change management and garner feedback.
- CRDC-H model will be stored in GitHub, to capture feedback and track changes.
- Documentation & changelogs for CRDC-H will be released and versions using GitHub
- Changes will be related via our newsletter, Slack, and GitHub.

Updates to the CRDC-H Harmonized Data Model

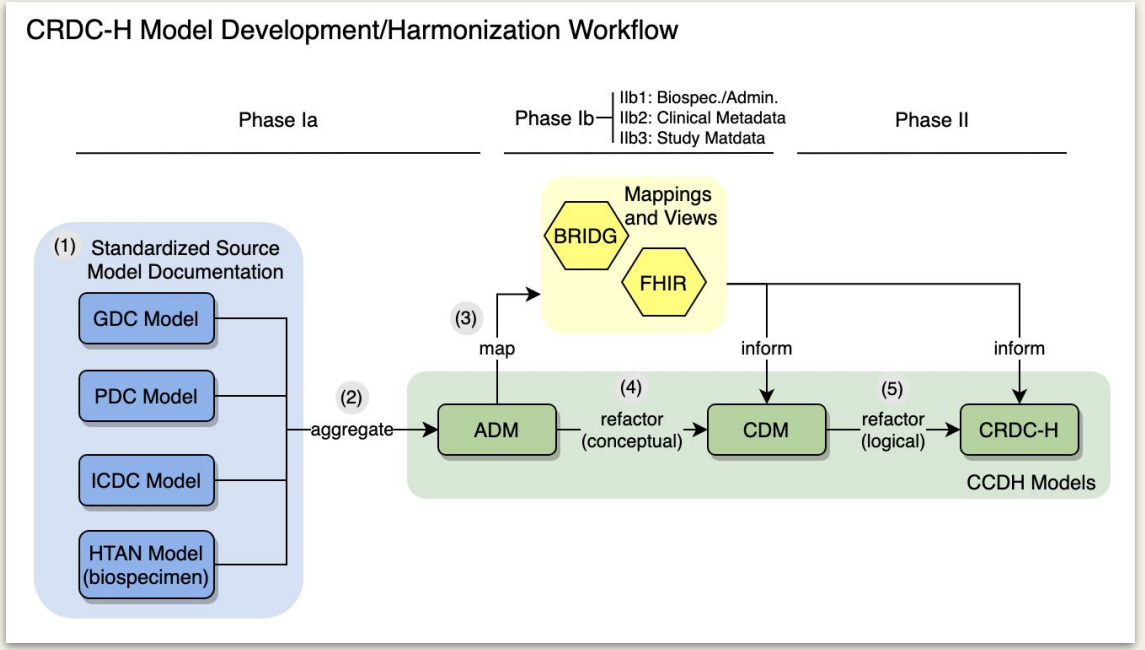
(Brian)



CENTER *for*
CANCER DATA
HARMONIZATION

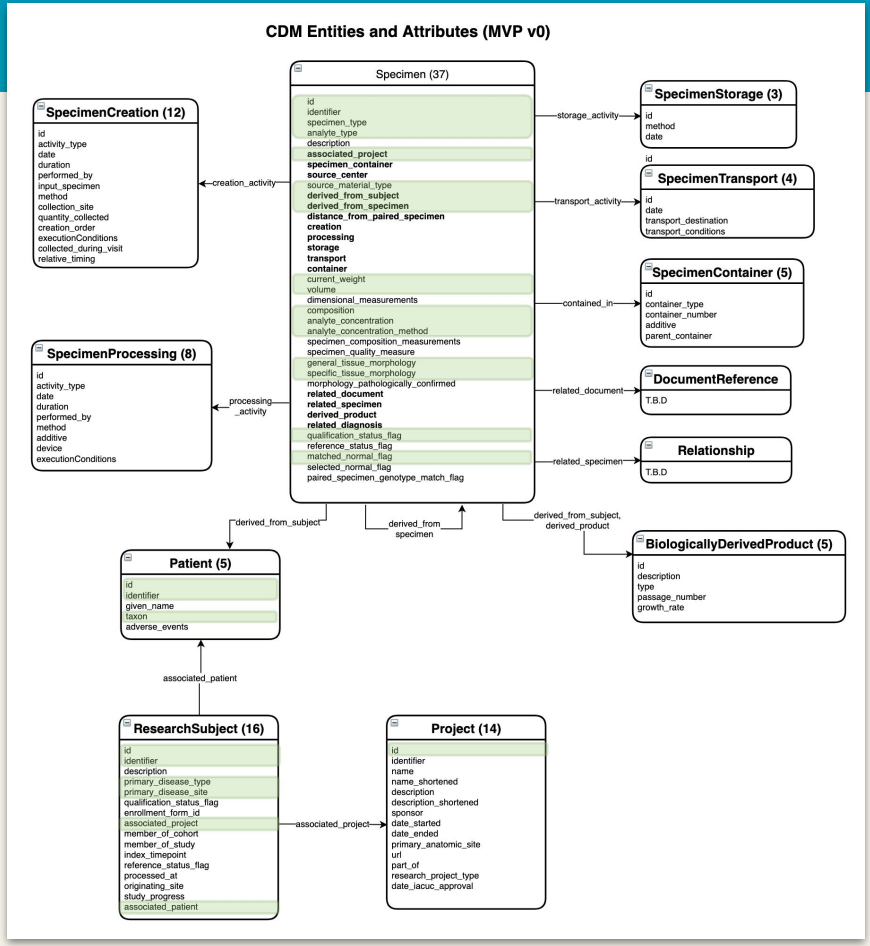
CRDC-H Model Development

- Model development process is iterative, but ***began with a high-level, coarse harmonization*** of source models into the ADM
- Subsequent modeling steps involve ***more nuanced refactoring of ADM into CDM***
- Finally this is represented in an implementable LinkML specification (CRDC-H)



CRDC-H Scope

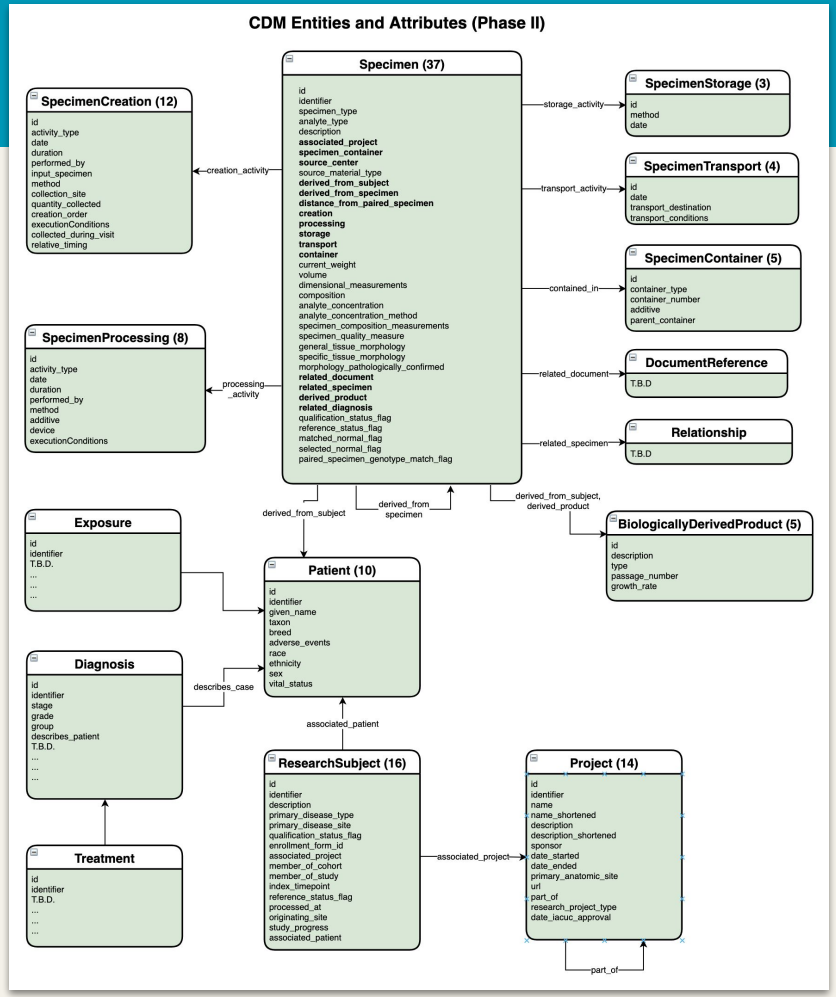
- MVP v0 release on Jan 26, 2021 represented *a subset of existing CDM entities and attributes*
  - Administrative and Biospecimen focused to align with the CDA pilot
  - Test tooling and model framework
    - Model generation from CDM to CRDC-H
    - Model documentation
  - Terminology bindings not present in model



CRDC-H Scope

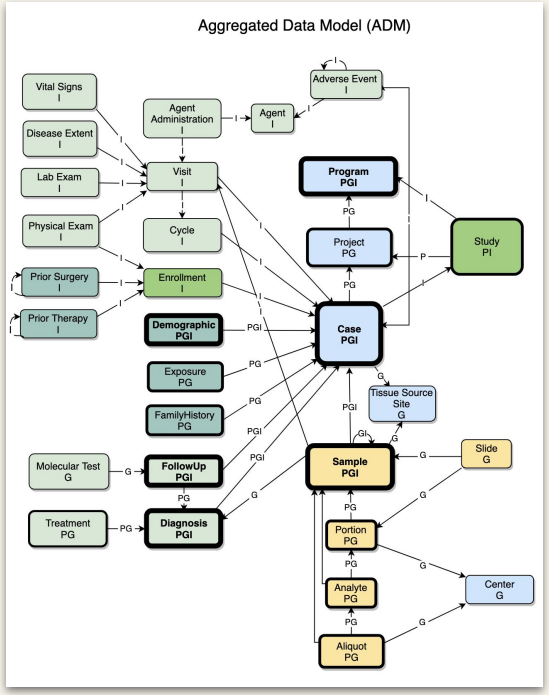
- End of Phase II release will include *remaining Biospecimen and Administrative subdomain entities*, along with *select Clinical subdomain entities*
  - Demographics
  - Diagnosis
  - Treatment
  - Exposure

- Terminology bindings will be included



CRDC-H Scope for Phase III Q1

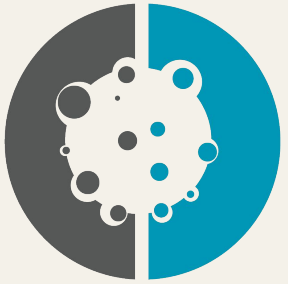
- Complete modeling of Clinical Data subdomain (October 2021)



- Initial modeling of Experimental subdomain (October 2021)
  - Generalized experiment metadata

- Coordinate inclusion of high-impact attributes from the Imaging Data Commons (IDC) with the CDA to support query
  - Faceted attributes in IDC Portal likely initial candidates

Terminological Alignment & Services for CRDC Nodes (Harold)



CENTER *for*
CANCER DATA
HARMONIZATION

Terminology Bindings Model – LinkML

```
enums:
  lens_color:
    description: Harmonized stoplight lens color
    code_set: ORS:pato_colors
    pv_formula: URI

  id_scheme:
    description: Stoplight identification schemes
    permissible_values:
      boise:
        description: Boise Identification scheme
      tf:
        description: Twin Falls Identification scheme
      poc:
        description: Pocatello Identification scheme
```

```
{ "members": [
    {
      "uri": "http://purl.obolibrary.org/obo/PATO_0000322",
      "code": "PATO:0000322",
      "defined_in": "PATO",
      "designation": "red",
      "definition": "A color hue with high wavelength …"
    },
    {
      "uri": "http://purl.obolibrary.org/obo/PATO_0000323",
      "code": "PATO:0000323",
      "defined_in": "PATO",
      "designation": "white",
      "definition": "An achromatic color of maximum brightness;  …"
    },
    …
```

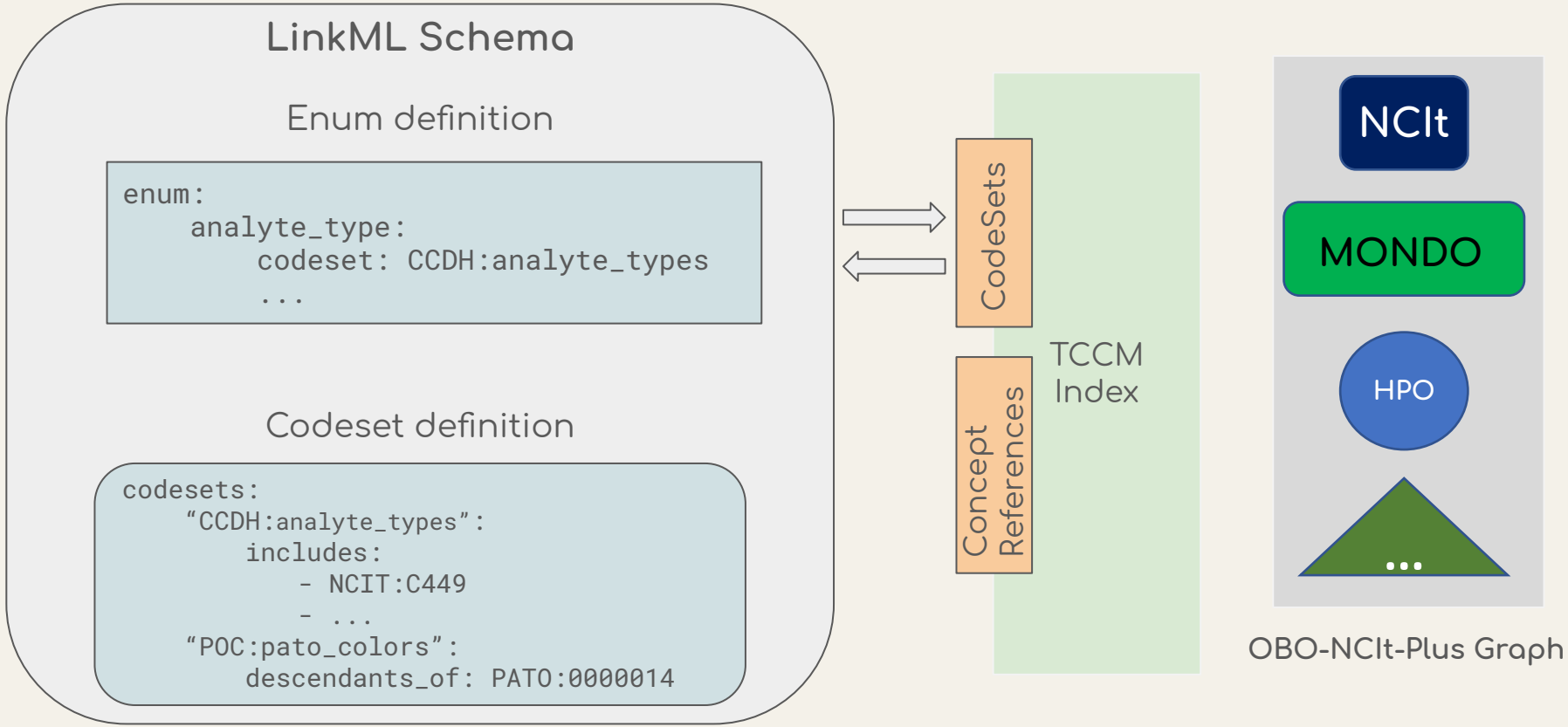
Equivalent to

```
lens_color_v3:
  description: Harmonized stoplight lens color URI
  code_set: ORS:pato_colors
  permissible_values:
    "http://purl.obolibrary.org/obo/PATO_0000322":
      meaning: PATO:0000322
    "http://purl.obolibrary.org/obo/PATO_0000323":
      meaning: PATO:0000323
```

Codeset based enumerations: flexible but semantically defined

LinkML: A developer and domain-friendly data modeling language that is transformable into common schema languages

Terminology Services – TCCM (Terminology Common Core Model)

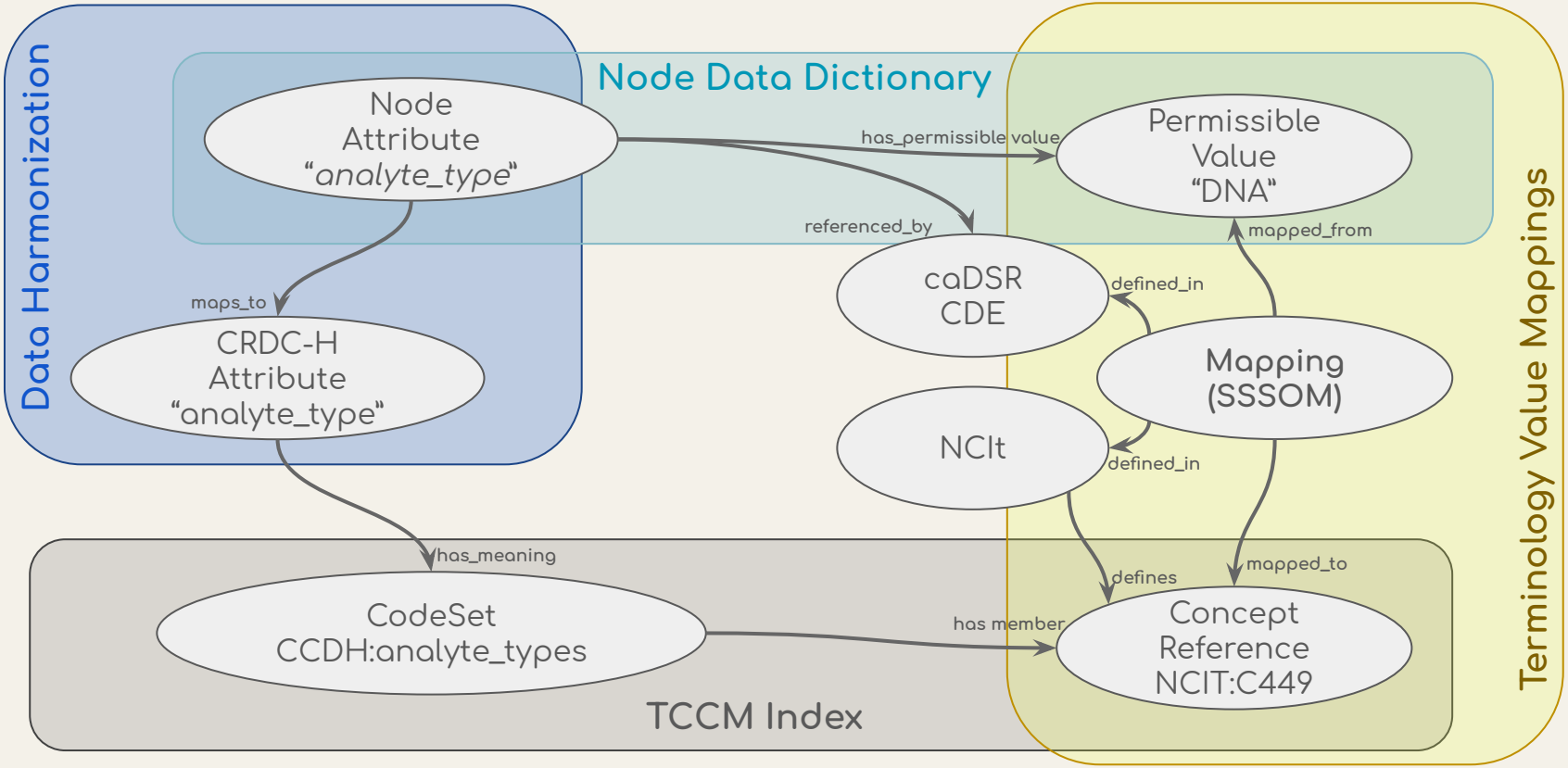


Value Mappings

- GDC/PDC value harmonization
  - Continue to map the permissible values in GDC and PDC to NCIt codes
    - Close to completion of mapping of values in specimen attributes
    - Mapped certain case-related attributes such as disease types
  - Discussed the strategies for next steps after the mapping
    - Determine the percentage of pre-coordinated values
    - For certain attributes, semantically integrate the bindings to domain ontologies such as MONDO using the OBO NCIt-Plus graph

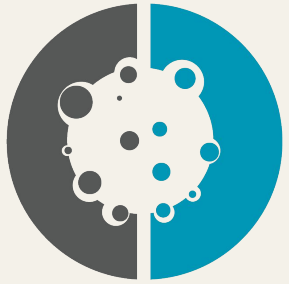
- Examining the data from GDC and PDC
  - Analyzed certain fields such as the analyte type, analyte type id, and their correlations in the analyte and aliquot data

Value Mappings Graph Model



Tools to put CRDC-H into use

(Jim)



CENTER *for*
CANCER DATA
HARMONIZATION

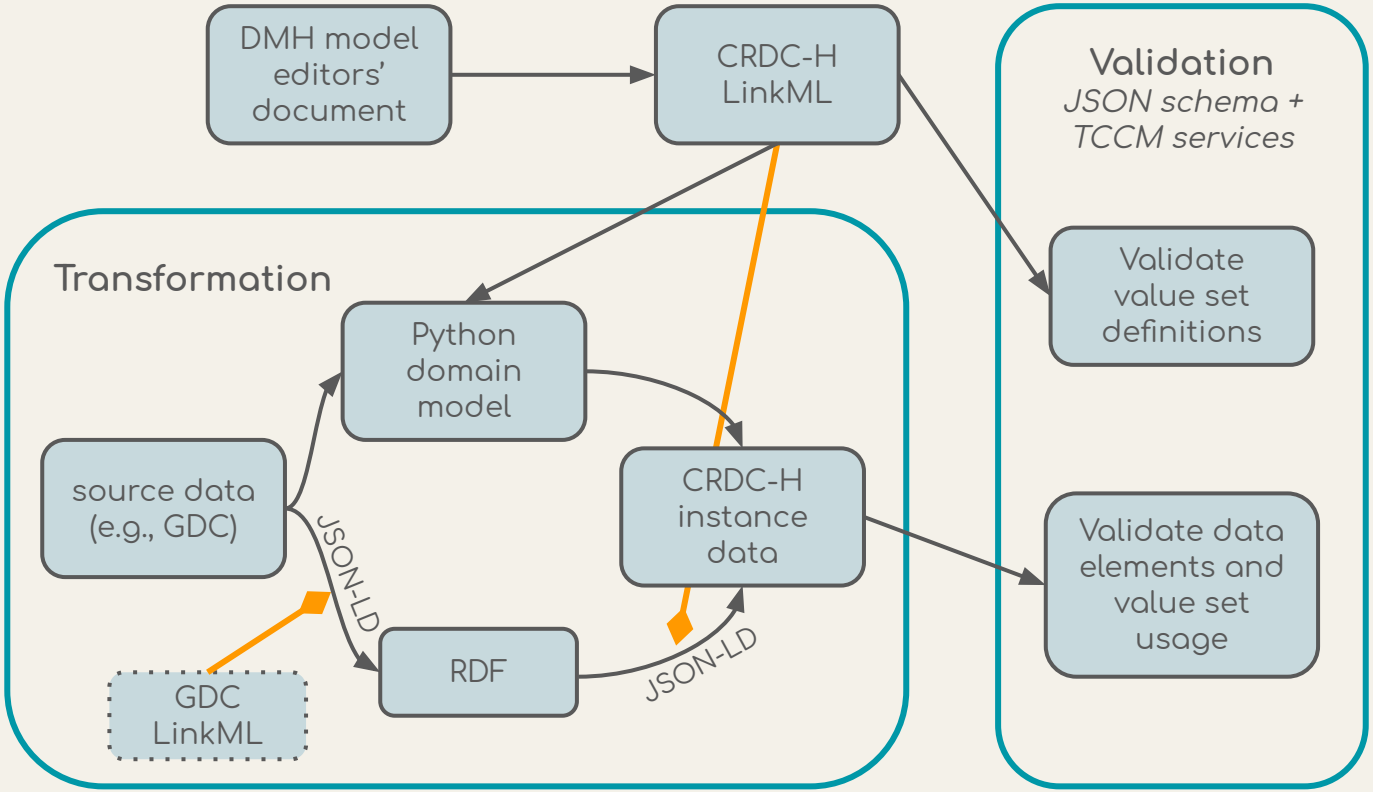
NCIt-Plus workflow

- The NCIt-Plus workflow integrates external ontologies into an NCIt-based unified ontology that can serve harmonized terminological content
  - initially prototyped with earlier funding
- Based on a set of term mappings (e.g., Uberon anatomy, CL cell types) and OWL axiom rewiring
- Recent work:
  - revamped the workflow to use the SSSOM mapping format, which is becoming a standard for sharing cross-terminology maps
  - retained links to original NCIt terms when replaced by external terms
- Plan to expand set of integrated mappings
  - e.g., GO processes (dependent on terminology harmonization needs)
- Working on browser interface for NCIt-Plus using OLS

Transformation and validation needs

- **Data in the nodes** currently use node-specific models that will need to be transformed into **CRDC-H instance data**.
  - CDA currently uses a mapping YAML file and custom code to do this transformation for their MVP; LinkML provides support for standardized approaches to ingest into CRDC-H.
- **Validate the CRDC-H model** to ensure that it is valid LinkML and that all the permissible values are correctly mapped.
- **Validate the CRDC-H instance data** to ensure that the data is represented correctly.

Transformation and validation tools



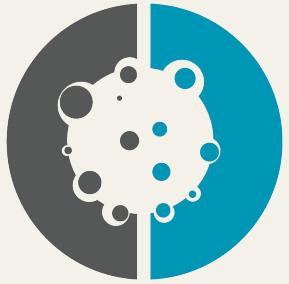
Indicates that the framework generates JSON-LD support

Looking forward

Harmonization services: *Enable tools like Ptolemy or MVS to utilize CRDC-H+TCCM*

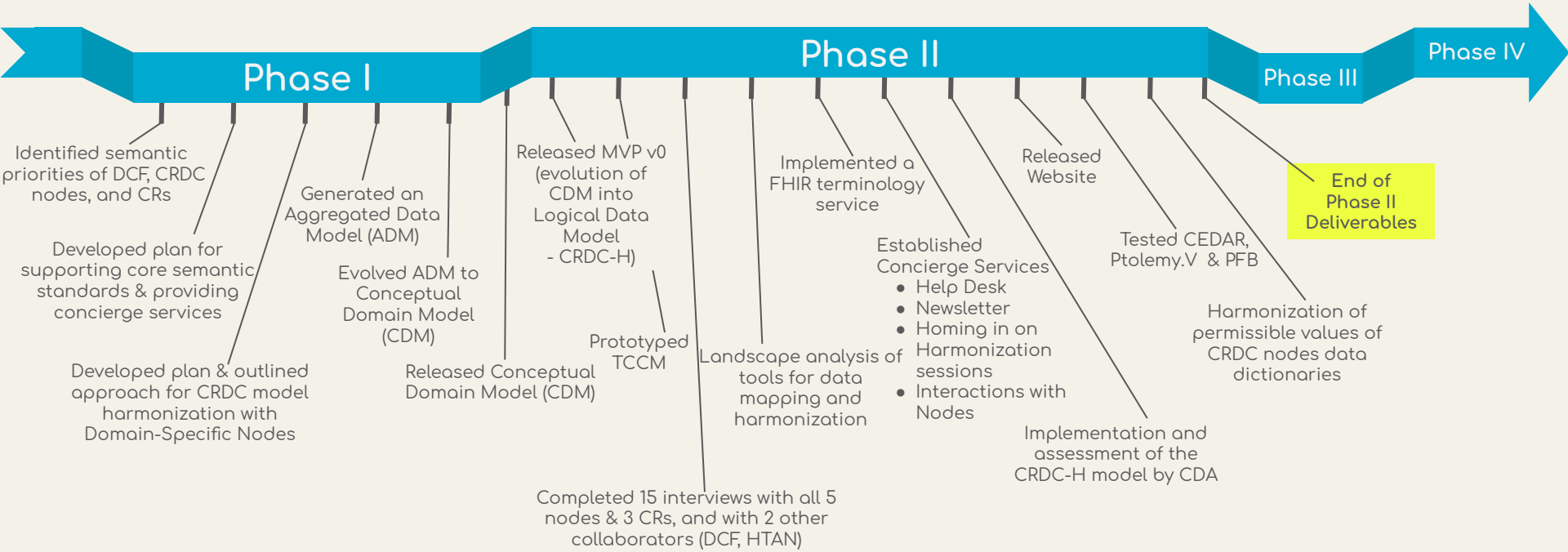
Standard validation workflows *for TSV, PFB*

Timeline Update & Summary of Planned Phase II Deliverables (Moni)



CENTER *for*
CANCER DATA
HARMONIZATION

Timeline


Phase I
Phase II
Phase III
Phase IV
Identified semantic priorities of DCF, CRDC nodes, and CRs
Developed plan for supporting core semantic standards & providing concierge services
Developed plan & outlined approach for CRDC model harmonization with Domain-Specific Nodes
Generated an Aggregated Data Model (ADM)
Evolved ADM to Conceptual Domain Model (CDM)
Released Conceptual Domain Model (CDM)
Released MVP v0 (evolution of CDM into Logical Data Model - CRDC-H)
Prototyped TCCM
Completed 15 interviews with all 5 nodes & 3 CRs, and with 2 other collaborators (DCF, HTAN)
Landscape analysis of tools for data mapping and harmonization
Implemented a FHIR terminology service
Established Concierge Services
Help Desk
Newsletter
Homing in on Harmonization sessions
Interactions with Nodes
Implementation and assessment of the CRDC-H model by CDA
Released Website
Tested CEDAR, Ptolemy.V & PFB
Harmonization of permissible values of CRDC nodes data dictionaries
End of Phase II Deliverables

Summary of planned deliverables End of Phase II (May 31st)

- Computable CRDC-H release
  - Biospecimens, administrative, clinical domains
  - LinkML computable specification and transformations into common schema languages
  - Deployed CRDC-H browsing user interface
- Terminology services
  - Harmonized NCIt-Plus graph
  - Terminology services for look up, equivalency
  - Terminology (NCIt-Plus) browsing interface using OLS
  - Terminology Common Core Model (TCCM) code set services
  - Graph driven REST API to browse mappings of nodes' attributes in their data model to the CRDC-H model and terminology mappings
- Tools for harmonization
  - Transformation toolkit to ingest data into CRDC-H
  - Validation tools for CRDC-H, terminologies, and codesets
- Community development
  - End-to-end inventory and recommendations
  - User Support and Engagement Plan
  - Community training & education sessions and documentation
  - Help desk & Website

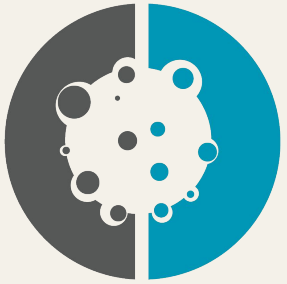
CENTER *for*
CANCER DATA
HARMONIZATION

Coming up for Phase III *... following our May release*

- Help users understand what harmonization looks like through visualization tools, documentation, & other concierge services
- Continue use of a transparent process to capture feedback, requesting new content, and enabling community discussions
- Terminology mapping tools
- Enhance validation toolkit for metadata validation and metadata mapping & transformation
- Complete modeling of Clinical Data subdomain and initial modeling of Experimental subdomain
- Coordinate inclusion of high-impact attributes from the Imaging Data Commons (IDC) with the CDA to support query
- Documentation to facilitate CRDC-H adoption and implementation
- Collaborate with cloud resources to have them evaluate the CRDC-H and provide granular and iterative feedback

Towards an understanding of the CCDH in the CRDC architecture

(Sam)



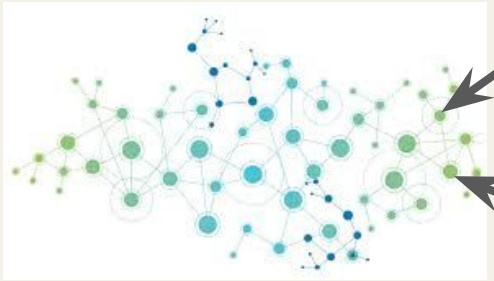
CENTER *for*
CANCER DATA
HARMONIZATION

Querying for data

Terminology Services & Validation Toolkit

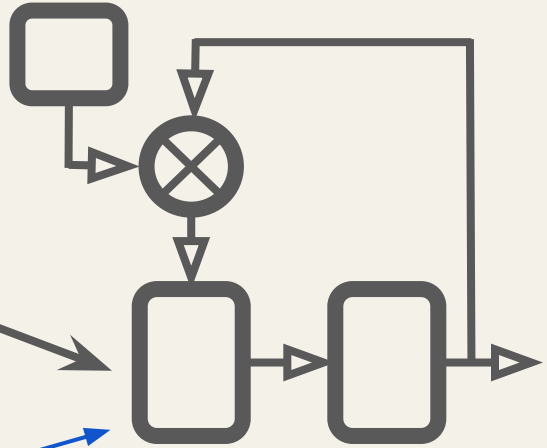
- Term lookup (including synonym, ID)
- Value set lookup & validation
- Equivalency determination
- Precompose vs. postcompose axiom validation

Semantically integrated terminologies & codesets



OBO-NCIt-Plus graph & Terminology Common Core Model (TCCM)

CRDC-H data model

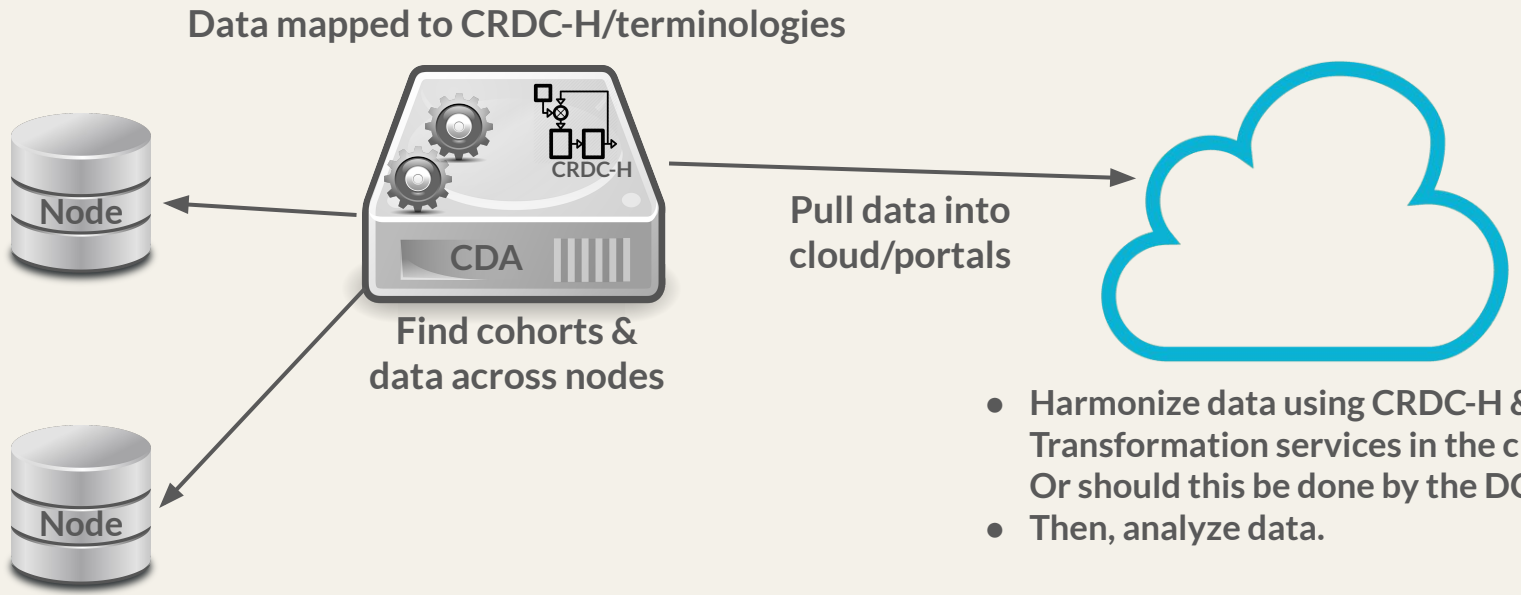


Ingest, encode, or migrate forward data using terminology & transformation services

Generate node data via a common CRDC-H API?

Utilize CRDC-H natively or map

Getting the data, harmonizing it, using it



- Harmonize data using CRDC-H & Terminology / Transformation services in the cloud resources? Or should this be done by the DCC or CCDH?
- Then, analyze data.

- Where and how will the data be harmonized for analytics?
- What the information architecture looks like will inform CCDH tool and services development.

End-to-end Requirements Analysis: Formal Recommendations

Report: bit.ly/crdc-e2e

- Broaden and augment CRDC governance structure.
- Create a strategic plan for the Cancer Data Ecosystem that includes governance, interoperability, communication, and shared resources.
- The CRDC-H (harmonized data model) should take precedence over any other dictionary or model in the CRDC for the higher level entities that it covers.
- A centralized submission process will greatly aid the ways in which data are consumed into the CRDC, especially for multi-modal data.

End-to-end Requirements Analysis: Formal Recommendations (continued)

Report: bit.ly/crdc-e2e

- Clinical data should have a single home within the CRDC.
- The CRDC needs to prepare for the consumption of real world evidence.
- Identifier management should be centrally managed and provisioned.
- Invest in information architecture.
- Consider technology management and technical architecture that promotes integration and coordination.

Acknowledgments

CRDC Nodes
CDS: Cancer Data Services
GDC: Genomic Data Commons
ICDC: Integrated Canine Data Commons
IDC: Imaging Data Commons
PDC: Proteomics Data Commons

DCF: Data Commons Framework - Infrastructure

Collaborators
Broad Institute FireCloud
CIDC: Cancer Immunology Data Commons
Gabriella Miller Kids First Data Resource Center
HTAN: Human Tumor Atlas Network
ISB: Institute for Systems Biology (Cloud)

Center for Biomedical Informatics & Information Technology

- Gilberto Fragoso
- Sherri de Coronado
- Melissa Cook
- Denise Warzel
- Cristina Russo
- Erika Kim
- Allen Dearry

Frederick National Laboratory for Cancer Research

- Todd Pihl
- Resham Kulkarni
- Mark Jensen

Samvit Solutions

- Smita Hastak
- Wendy Ver Hoef
- Charles Yaghmour

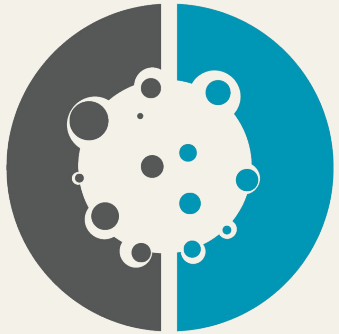
Cancer Data Aggregator

- Kathy Reinold
- Annie Kuan
- Brian O'Connor
- Alex Baumann
- David Pot
- Jack DiGiovanna

CENTER *for* CANCER DATA HARMONIZATION

Questions & Answers

These slides:
bit.ly/ccdh-q1-2021



CENTER *for*
CANCER DATA
HARMONIZATION

CCDH Quarterly Report to Federal Oversight (NIH/NCI and FNL)
Date: January 13, 2021

ccdh.cancer.gov