Colon CFR
Development of an Enhanced Biospecimen Information Technology Infrastructure

Background and Rationale

The Colon Cancer Family Registry (C-CFR) centers collect a number of biospecimen types, including blood, buccal cell, tumor paraffin block, and fresh frozen tissue. From these, the centers derive additional materials, such as DNA samples, immortalized lymphocyte cell lines, and pathology slides. Many of these biospecimens are consumed during routine laboratory core tests specified in C-CFR protocols. Many specimens are also banked in long term storage and are available to collaborating investigators.

The principal data dissemination point for the C-CFR is the Informatics Support Center (ISC). Although data on biospecimens are regularly transmitted to the ISC along with questionnaire data and laboratory testing results, biospecimen data have thus far been disseminated primarily from the individual centers. To a large degree this is due to the fact that the standardization of transmitted biospecimen data fields is incomplete. The original informatics center biospecimens database was insufficiently flexible to accurately reflect the specimens at the sites. For example, it was not able to capture multiple types of specimens (normal, adenoma, cancer) that existed in the same surgical specimen. This information was carefully traced within the local databases but capturing this level of detail in a consistent manner across the C-CFR was not accomplished in the central database, and subsequently mapping these specimens to test results (MSI, immunohistochemistry) and transmitting this data lead to unclear aggregate data. Fortunately, the data at the sites was interpretable locally, so current studies have received the correct biospecimens and information and the current informatics system is rewriting the Biospecimens and Dispatching data dictionaries.

Although biospecimen protocols were standardized early in the history of the C-CFR, the laboratories affiliated with most centers were already using existing LIMS (laboratory information management systems) or analogous technology and continued using these existing systems for C-CFR biospecimen tracking. Local data are transformed from these systems into a common data model (CDM) for transmission. The first generation CDM, which is still in use, allows for heterogeneity among these systems to persist into the central database.

Now that the transition of the ISC from the University of California, Irvine to Research Triangle Institute (RTI) has been achieved, RTI is spearheading a major effort to complete the centralization and standardization of biospecimen tracking information so that a single source of information on available C-CFR biospecimens can be realized. This effort, which has recently begun, is being undertaken collaboratively with the

Biospecimen Working Group (BWG). The major objectives are quick and easy access to the following information across centers:

- The set of specimens meeting a given set of person-level or specimen-level criteria
- The current inventories (i.e., amounts remaining) of specimens
- Critical events, parameters, and protocols in the processing of specimens
- The history of biospecimen distribution

This information would be valuable to a number of stakeholders: It would help investigators plan studies. It would help the Advisory Committee (AC) make decisions about access to specimens for proposed studies. It would help trigger the replenishment of renewable biospecimen types. It would identify specific biospecimens of interest to investigators. It would provide valuable usage statistics to NCI.

The enhanced information technology infrastructure, which we refer to as the Biospecimen Tracking System (BTS), will be built on a foundation of rigorously-defined data fields composing an enhanced CDM. Data transmitted from the centers in the form of these fields will be stored in a relational database management system such that they are integrated via anonymous linking identifiers with the pathology, epidemiology, diet, family history, and molecular data that are captured by the C-CFR. (This is the current architecture, although the corresponding database tables will be modified and enhanced as required.) The system will also provide specific user interfaces for different sets of stakeholders.

Fortunately, many issues related to the management of biospecimen data are universal, and well-designed approaches exist. We will draw from two sources of design knowledge in the crafting of this system: First, the document *First Generation Guidelines for NCI-Supported Biorepositories* will be a major source of system requirements. Second, we will leverage IT resources developed by NCI Center for Bioinformatics. These will include common data elements (CDE) for defining biospecimen data fields and technologies for systems integration developed through the Cancer Biomedical Informatics Grid (caBIG) project.

Here we describe the system as currently envisioned along with our plans for development. Throughout we describe work accomplished to date.

Developing the Enhanced CDM

The current CDM, developed by the previous ISC contractor, contains fields for recording attributes of paraffin blocks, slides, lymphocytes, buccal cell, and DNA specimens collected. It also contains fields for tracking dispatches of biospecimens. An entity-relationship diagram depicting data model is provided in Figure 1. The model was developed prior to the commencement of collaborative studies that included biospecimens.

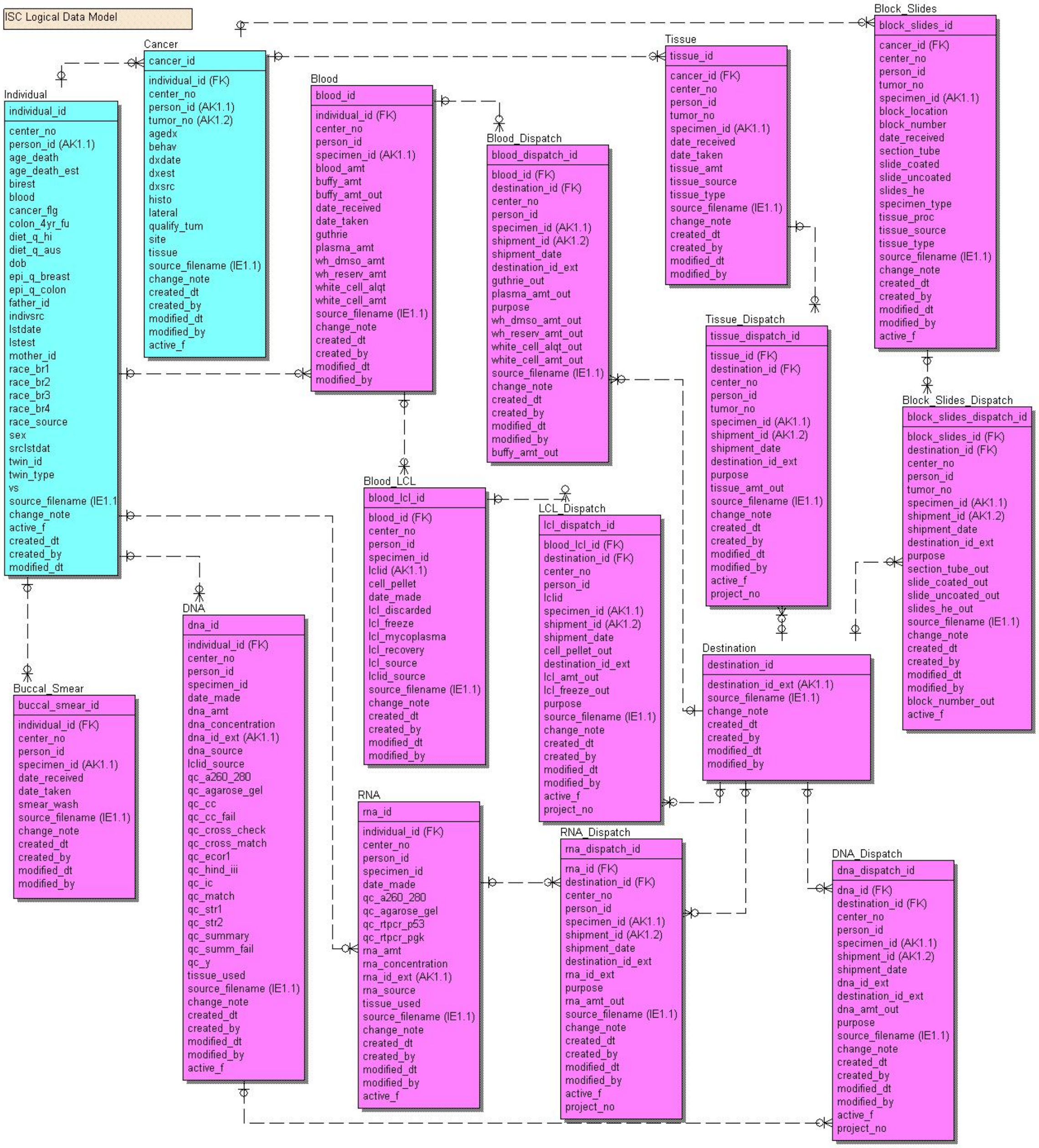


Figure 1. Entity-Relationship diagram depicting the current common C-CFR biospecimen data model.

RTI and members of the BWG have recently begun reviewing the field definitions within this model (available from the ISC web site http://www.cfrisc.org) in light of the needs of the various stakeholders and with the intent of discovering and normalizing sources of variation across centers. This review is also examining how the data model supports the requirements specified in the NCI biorepository guidelines. To date, several disparate issues have been identified. One of the major sources of variation across centers is the use of different units of measurement. For instance, the units of measurement for blood draw size is *vials*. However, the capacity of a vial differs across centers. The response to such cases has been to tighten the corresponding data field definitions to remove such ambiguity.

To aid in the development of the data model, a minimal information specification has been developed. From this, a checklist was developed by Robert Haile to ascertain how individual local tracking systems support this specification. This checklist is provided in Appendix A. The results from individual centers are provided as separate documents that are included with this document. Most centers are substantially in compliance with the specification. For centers not in compliance, USC has offered to make their system available if centers feel the effort of upgrading their local system is too great. RTI would help convert existing local data into the USC data model so that historical data would be available through the USC system.

Another major objective of this review is to harmonize the CDM with caBIG-supported common data elements (CDE). To assist in this effort, we will work with staff from NCI Center for Bioinformatics (NCICB) familiar with biospecimen CDEs. We have already had several discussions with NCICB staff to talk about general issues and strategies.

The enhanced CDM will be published on the ISC web site. The ISC will work individually with centers to ensure they understand how to map local schema to the common data model. The ISC will then review submitted data to ensure that data fields are being used correctly and consistently across centers. For any deviations found, the specifications for the corresponding fields may be modified to provide greater clarity.

Note that if the concept of the “split” biospecimen repository (see accompanying document) is realized, no major modifications to this approach should be necessary. The central repository essentially becomes another center. One major concern will be to ensure that proper linking identifiers are used so that data on biospecimens stored at the central repository can be properly linked back to the other C-CFR data from the originating centers.

Finally, the C-CFR has had preliminary discussions with Dr. Carolyn Compton of the OBRR whose office will serve as advisors to the C-CFR and to RTI as these aforementioned enhancements are implemented.

Developing User Interfaces

To date, no specific user interfaces for accessing biospecimen data within the central database have been successfully developed. The original informatics center did attempt to develop an OLAP cube-based interface, which was beta-tested by some of the CFR PIs, but was never finalized. With the transfer of the informatics center to RTI, development of his type of functionality has begun all over again.

Currently data are obtained directly from the database management system, which limits access to ISC staff members who have database accounts. Web-based user interfaces can provide a level of data access to users from outside of the ISC. There are at least seven major classes of stakeholders, for whom different interfaces to the system will be provided. These classes along with the corresponding levels of access are summarized in Table 1.

User Class	Level of System Access
ISC staff member	Individual records, all data fields
Center (non-ISC) staff member	Individual records for affiliated center only, all data fields
Potential collaborating investigator	Summary data only
Collaborating investigator	Records and fields meeting approved inclusion criteria
Biospecimen Working Group	Summary data plus specimen dispatch request form
Advisory Committee	Summary data only
General Public	Canned summary tables only

Table 1. Identified classes of users of the biospecimen data access system with corresponding level of access.

ISC staff members need access to all data for filling data requests, ensuring data quality, and other activities. Their interface would provide access to all records and all data fields. Center staff members would have a similar level of access, but only to the subset of data from their center. Potential collaborating investigators would benefit from access to summary data in developing their study designs. These data would consist of numbers of specimens meeting certain criteria along with distributions of specimen amounts currently in the inventory. (Access to record level data is restricted to investigators with approved studies—i.e., collaborating investigators). Collaborating investigators will have access to records and fields approved in their research application. The BWG and AWG members would benefit from summary data on the number of specimens available for a particular proposed study. Additionally, the system will provide the BWG with a means for requesting dispatch of specific specimens meeting an approved study's inclusion criteria. Finally, pre-generated (i.e., "canned") tables of basic summary data will be made available to the general public.

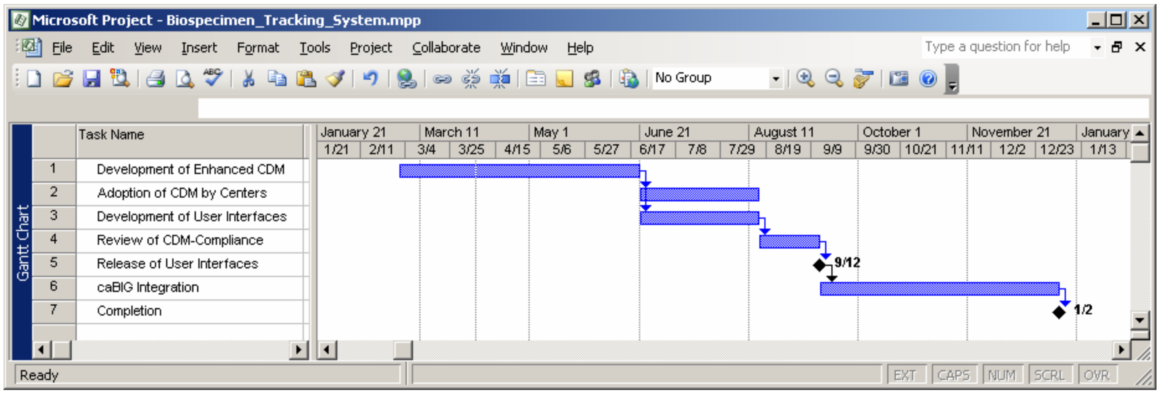
The ISC will provide these user interfaces through its web portal. Currently the ISC is developing general purpose "query wizard" technology, which will be released in late April, 2007. This technology will enable the user to build queries for data by selecting data fields and setting conditions on those fields. This technology will be employed to provide these user interfaces. For user classes with record-level access, the query result will consist of individual records. For user classes with summary-level access, the result will consist of record counts and distributions.

Developing Systems Integration Points

NCI has realized that the potential utility of data can extend far beyond the bounds of the study under which they were collected. To foster a greater degree of data sharing and integration across studies, NCI has supported the development of infrastructure for integrating data and systems. One of the major thrusts of this effort is the caBIG initiative. The ISC has been charged with providing caBIG-compatible interfaces. Such interfaces would facilitate the analysis of these data with third-party software tools and would enable integration with other databases. The harmonization of the CDM with CDEs described above is one step towards caBIG compatibility. A major technological issue to work out before this is implemented is how to provide appropriate access control to data. The level of access through caBIG should mirror that available through the web user interfaces described above. The mechanism of integration would be a suitable web service that is caBIG silver-level compliant.

Development Schedule

The current schedule for development of the BTS is shown in Figure 2. Development of the enhanced CDM began in March, 2007 and is anticipated to conclude in mid-June. After completion of this model, the centers will work on compliance while the ISC develops user interfaces based on this model. After these activities, the ISC will conduct a review of center compliance with the CDM. After this, the user interfaces will be released to the production web portal. The current projected timeframe for this milestone is mid-September, 2007. After release of the interfaces, the ISC will work on caBIG compliance, completing this step in early January, 2008.



	Task Name
1	Development of Enhanced CDM
2	Adoption of CDM by Centers
3	Development of User Interfaces
4	Review of CDM-Compliance
5	Release of User Interfaces
6	caBIG Integration
7	Completion

Figure 2. Gantt chart depicting schedule for biospecimen tracking system development.

Hawaii Colon-CFR Biorepository Database

Database Overview

The Hawaii C-CFR biorepository database system was developed in Microsoft Visual FoxPro on a Novell network. The system manages the collection, processing, storage, retrieval and dispatch of all biospecimens (blood, buccal cells, PET and DNA) and maintains an ongoing inventory of all samples, including type of specimens, number of aliquots and volume/amount. System and data access is restricted to authorized users at the database and network levels.

Database Features

Blood: Collection & Processing.

Blood Collection

Most blood draws are done by the study's medical technologists. For subjects living on neighbor islands or the US mainland, blood draws are scheduled at clinical labs. A manual log is used to record the date and time of collection, date and time of processing, and date and time of storage as well as visual observations regarding the status of each blood draw. Aliquot label details (e.g. P-01, P-02, …) and barcodes are also recorded on the log.

☑ Record blood collection date and time

☑ For remote draws, record date and time specimen is received

☑ Assign a unique 'Spec_no' for each blood specimen collected

This ID is applied to subsequent blood components: Blood Spots, Plasma, Buffy Coat, Lymphocytes, LCL and DNA.

☑ Track the number of separate blood draws received from each patient

☑ Record an ancestry questionnaire for each subject drawn

Blood Processing

☑ Record id of technician who processed blood

☑ Record volumes of EDTA and ACD blood collected

☑ Record dates of EDTA and ACD blood processing

☑ Record dates of EDTA and ACD blood storage

☑ Assign each aliquot a unique barcode

☑ Assign each aliquot a unique storage location that identifies its freezer, shelf, rack, box and box position

☑ Record in a memo field the condition of blood received (e.g. short draws, hemolyzed, etc.) and processing outcome

Paraffin Embedded Tissue (PET): Collection & Processing

☑ Record hospital, pathology number and pathology date for PET blocks and/or slides

☑ Record hospital assigned block number for all PET blocks collected

☑ Record date of receipt for PET blocks and/or slides collected

☑ Assign a unique 'Spec_no' for each PET block collected

This ID is applied to subsequent block components: Slides, paraffin sections and DNA and links these to the corresponding C-CFR pathology form and other CFR data.

☑ Assign each sample (block, slide, section for DNA) a unique storage location that identifies its freezer or cabinet, shelf, rack, box, and box position.

DNA: Extraction & Storage

The Genomics Shared Resource (GSR) lab of the Cancer Research Center of Hawaii extracts and stores DNA for this study. The lab assigns its own unique identifier to each subject. DNA samples are tracked by this GSR ID. Samples dispatched in vials are labeled with this identifier.

The existing DNA extract database is part of the Hawaii C-CFR tracking system. DNA dilutions and dispatches are tracked in Microsoft Excel. Dispatch spreadsheets are imported into Visual FoxPro tables. A new database system is being developed for the GSR lab using Microsoft SQL Server and Visual Basic.

☑ Record date of DNA extraction

☑ Record the source product and its barcode

The barcode serves to connect the DNA to the source product 'Spec_no'.

☑ Record DNA extract specifications: concentration, volume and amount

☑ Assign each sample a unique storage location that identifies its freezer, shelf, rack, box and box position.

Biospecimen (Blood, Tissue and DNA) Dispatch

☑ Record dispatch date, destination and purpose

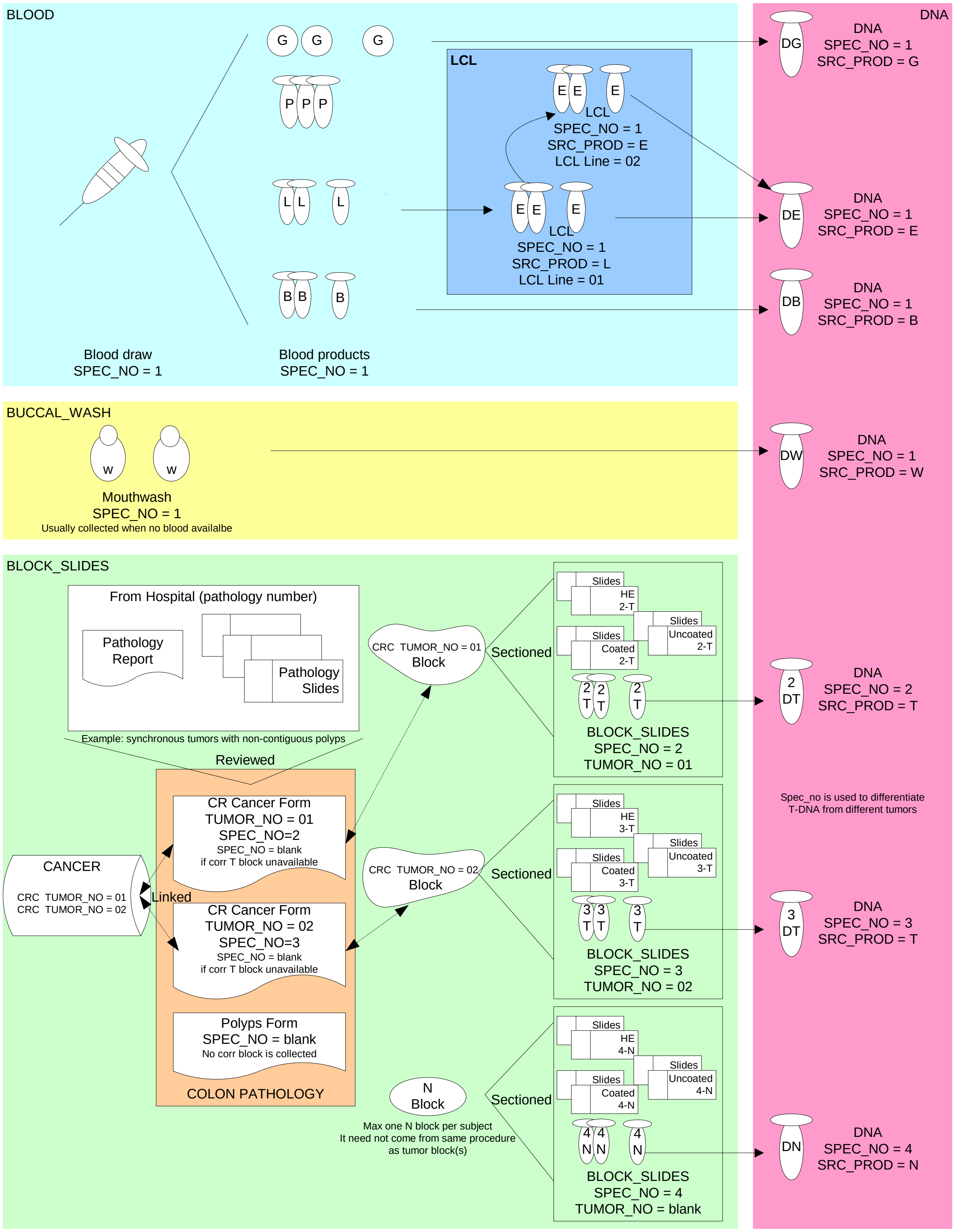
☑ Record type and number of aliquots dispatched

☑ Record ID numbers of samples dispatched

☑ Record amount of sample dispatched

☑ Update biorepository database to reflect remaining types and numbers of aliquots and volume of samples.

Hawaii Biospecimen Flow Diagram



Mayo C-CFR Biorepository Database
Biospecimens Accessioning and Processing (BAP) Shared Resource

The BAP Laboratory: The Biospecimens Accessioning and Processing (BAP) Shared Resource was created to meet the evolving needs of the Mayo Clinic Cancer Center (MCCC), the Mayo Advanced Genomics Technology Center (AGTC), the Center for Individualized Medicine (CIM), and other Mayo researchers for specimen acquisition and processing. The roots of the BAP facility lie in the clinical laboratory, specifically, the Molecular Genetics Laboratory (MGL) in the Division of Laboratory Genetics, Department of Laboratory Medicine and Pathology (DLMP). As nucleic acid extraction is an integral part of molecular genetic testing; and, since the MGL had developed the appropriate technology and methods base for clinical use, it was logical that this laboratory also serve the accessioning and nucleic acid extraction needs of the research community at Mayo. Thus, research samples were handled alongside clinical specimens for over 15 years. In 2002, with the formation of the Mayo Genomics Initiative and the ATGC, the research accessioning and processing activities split from the MGL and merged with the EBV transformation/cell immortalization shared resources to form BAP. BAP was then able to focus on a mission of support of the research community.

The BAP Shared Resource is integrated with the Microarray, DNA Sequencing, Genotyping, and Cytogenetics Shared Resources to form the Genomics Resource Center on the 13th floor of the Stabile building. The Genomics Resource Center is supported collaboratively by both the MCCC and the Mayo Genomics Initiative to provide institution-wide services for biomedical genomic research. This arrangement minimizes duplication of services within the institution and maximizes efficiency, quality, and availability of services for Mayo investigators. In addition, this organization is designed for better logistical flow of specimens and will maximize specimen information integration into a single database (Research Accessioning and Tracking System [RATS]) electronically linked to the Mayo Life Sciences Warehouse database.

Currently, the BAP Shared Resource employees 13.8 full time equivalents (FTE) plus a full-time supervisor and assistant supervisor. 10.6 of these FTE are employed in the nucleic acid extraction facility. The nucleic acid extraction facility occupies a 1750 square foot laboratory that is equipped as a state of the art extraction core. Large equipment includes a Gentra AutoPure DNA extraction robot, AutoGen FlexStar DNA extraction robot, a Beckman Coulter Biomek NX S8 Liquid Handler, a Beckman Coulter DTX880 microplate spectrophotometer, and an Autogen L245P RNA extraction robot. Fully validated protocols are currently in use for specimen accessioning, DNA extraction (using manual and automated platforms and using non-organic or phenol-chloroform chemistries), RNA extraction, DNA and RNA purity assessment (spectrophotometry and Agilent capillary electrophoresis, respectively), and adjustment of nucleic acid to standard concentrations. In addition, the laboratory has a fully developed quality control and quality assurance program, reflecting the roots of the facility in the clinical laboratory environment. The laboratory is overseen by two Co-Directors, Drs. Wilma Lingle and W. Edward Highsmith. The nucleic acid extraction facility is currently handling 58,000 samples per year under a wide variety of protocol types.

The services provided by BAP include the following:

1. **Biospecimen Acquisition and Electronic Accessioning.**
2. **Specimen Processing.**
    a. Isolation and freezing of buffy coats
    b. Plasma and serum separation and storage
    c. Dissection and flash or controlled temperature freezing of solid tissues
    d. Preparing frozen tissue, buffy coats, or cultured cells for nucleic acid extraction
    e. Sample retrieval
3. **Peripheral blood mononuclear cell cryopreservation and EBV immortalization.**
    a. Cryopreservation of peripheral blood mononuclear cell
    b. Preparation EBV-transformed B-lymphocyte cell lines
4. **Nucleic Acid Extraction.**
    a. DNA and RNA extraction from fresh and paraffin embedded material
    b. Quality assessment using spectrophotometry.
    c. Post extraction aliquoting and storage
    d. Sample retrieval

The RATS database. Currently, BAP samples are accessioned into the Mayo Genetics Systems (MGS) database. MGS is a Sybase based application running on DB2 that was designed, built, implemented, and is maintained by the Department of Laboratory Medicine and Pathology Information Technology (IT) liaisons. The system was designed as a clinical resulting and reporting tool, and has been used by the MGL, the MGL DNA extraction core, and later, BAP, for many years. However, MGS has some shortcomings as a research sample database and tracking tool. Thus, we are currently engaged in a collaborative effort with LabVantage, Inc. to develop a database specifically designed for high-capacity research sample accessioning and tracking. For samples originating at Mayo and collected under Mayo IRB protocols, RATS will draw in patient demographic data, associated pathology reports, and other items from the electronic medical record in accordance with the patient's informed consent status. A unique identifier is assigned to each specimen, and fields containing other identifiers that could compromise patient confidentiality are hidden to investigators using the database. Access to the database is restricted to registered users. Registered users have access only to specimen information collected under their particular IRB-approved protocols.

Samples that are received in the laboratory are associated to their respective IRB-approved protocol which has been predefined within the RATS database. The RATS database records sample movement, processing, creation, and consumption through all predefined processes that the individual sample travels within BAP. Processing pending lists are used to control sample grouping and movement throughout the laboratory. As a sample moves from one processing work list to another, billing events are created. Sample integrity checks are encountered throughout processing, and these are internally documented within the RATS database, and completion is required in order for the sample to complete a specific processing point. Each sample is given a unique storage location within a freezer. After the samples are in storage, post storage movement, i.e., withdrawal of sample by an approved investigator for an experiment, is monitored and tracked. This tracking of samples, including what type of sample was requested, how much of each sample was requested, the number of samples requested, and the destination of the biospecimens, is documented within a project management access database. Within the database, the record of the volume of sample remaining is adjusted based upon the specific amount requested.

At this writing, the LabVantage software has been received by the Center for Individualized Medicine's software management committee, the group charged with implementation of the RATS system. Multiple programmers from the Cancer Center Systems group, the Center of Individualized Medicine IT support group, and specialized project managers from the Advanced Genomics Technology Center are currently engaged in developing the RATS system. Over 80 complete use cases have been defined for individual tasks within BAP.

Database Features:
Blood: Collection, Processing & Dispatch

☑ MGS tracks and records daily delivery of blood

☑ Record the arrival time of blood

☑ Record the name of each technician working on blood processing

□ Record the completion time of blood processing

☑ Record 'Study_ID" and IRB Protocol number submitted by collection center

☑ Generate a unique 'Specimen_ID' for each blood received.

  ○ This ID is applied to subsequent blood components: Blood Spots, Plasma, BC/RBCs, Lymphocytes and purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample generated is given a unique 'Barcode_ID'

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer, shelf, rack, box and well

□ Record if sample is collected from heparin or EDTA tubes.

☑ Record the condition of blood received, and processing outcome.

□ Record the reasons for pending redraw request

☑ Track the number of blood draws received from each patient.

☑ A Project Request Database that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples requested

  - ○ Tied to Project Request Number

☑ Record amount of sample requested

□ Records confirmation of dispatch receipt

☑ Records start date of dispatch

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request.

□ DNA samples that are less than 2 ug/ tube will be flagged as “low inventory”

FHCRC C-CFR Biorepository Database (“Vidro”)

Database Overview:

This database was designed to accommodate the growing biorepository here at Seattle’s CORE study. This system is designed using Microsoft SQL Server & C#.

The advantages of this system are its generic design, allowing for storage of multiple types of specimens, and its de-normalized attribute system, which allows a variable number of data points to be stored for each specimen. So adding new data points is seamless.

This system can manage the collection of all biospecimens, catalog dispatches and maintain an ongoing inventory of each sample. It can also track variable amounts of data related to the specimen that can be used as criteria towards dispatches.

Database Features:

Blood: Collection, Processing & Dispatch

□ A Blood Log that tracks and records daily delivery of blood

☑ Record the arrival time of blood

☑ Record the completion time of blood processing

☑ Assign unique id that identifies participant that provided the sample (can relate to and index to other systems as required).

☑ Generate a unique id for each blood received.

  ○ This ID is applied to subsequent blood components: Blood Spots, Plasma, BC/RBCs, Lymphocytes and purified DNA. Each sample has a reference to another “sample” that may have fathered it. For example, a DNA sample derived from a buffy coat, will have the id of the buffy coat created it, within its attribute. Similarly that buffy coat will have the id of the vial of blood that produce the buffy coat.

☑ Each sample generated is given a unique ‘Barcode’ value.

☑ Each sample is given a unique ‘Storage_Location’ that identifies its freezer, shelf, rack, box and well

☑ Record if sample is collected from ACD or EDTA tubes

☑ Record the condition of blood received, both ACD and EDTA tubes, and processing outcome.

☑ Record the reasons for pending redraw request

☑ Track the number of blood draws received from each patient.

☑ A Shipment log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators.

☑ Record destination of biospecimens for dispatch

☑ Record basic information on the scientific study related to the dispatch

☑ Record multiple shipments related to that study.

☑ Record processing instructions for samples related to a dispatch.

☑ Search engine to find and assess samples for a dispatch from multiple sets of criteria.

☑ Record number of samples requested

  ○ Selected by searches, which mirror criteria for the study, a clipboard mechanism is used to manage complex searches that involve two or three tiered logic to discern eligible samples.

☑ Record amount of sample requested

☑ Records confirmation of dispatch receipt

☑ Records start date of dispatch

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request.

  ☑ DNA samples that are less than 2 ug/ tube will be flagged as “low inventory”

    o These flagged aliquots will not be selected for any future dispatches

Paraffin Embedded Tissue: Collection, Processing and Dispatch

* Note that this branch of biospecimen collection is currently managed using Microsoft Access and Microsoft Excel. These data will be migrated to the same SQL database currently used for blood.

☑ A PET Log that records receipt of PET blocks and/or slides

☑ Record the arrival time of PET blocks and/or slides

☑ Record the completion time for block processing

☑ Record ‘Study_ID” submitted by collection center

  ○ Synonymous to “Person_ID”

☑ Record ‘Pathology_ID’ given by hospital

☑ Record ‘Block_ID’ given by hospital

☑ Generate a unique ‘Specimen_ID’ for each PET block or batch of slides received.

  - ○ This ID is applied to subsequent block components: Slides, paraffin-ribbons, TMA cores and purified DNA. This is the linkage variable throughout this portion of the database.
  - ○ This ID can be matched to the hospital generated ‘Block_ID’

☑ Each sample is given a unique ‘Storage_Location’ that identifies its freezer or cabinet, shelf, rack, box, and well.

☑ A Dispatch Log that records biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples requested

☑ Record amount of sample requested

☑ Records confirmation of dispatch receipt

☑ Records start date of dispatch

☑ Records blind duplicates

☑ Record case/control pairs on a dispatch

☑ Upon completion of dispatch biospecimen inventory (number of slides, paraffin-ribbons, and TMA cores remaining) is updated based on dispatch request.

☑ Record association to pathology data for tissue samples

☑ Record basic tumor information for tissue samples derived from a tumor

Security & Backups:

The database uses integrated NT security, meaning it works directly with the FHCRC network to establish visibility into the data, allows users only access to the data procedurally, meaning direct access to the physical data tables is not permitted, so even if one were to gain access to the database through manipulation of the network, one could not see, view or even delete any of the data. One could only reference the data via the procedures the software allows one to view. The procedures ensure the data are safe and dangerous operations such as ‘deletes’ are not permitted. Deletions are simulated with ‘logical deletions’, but the record itself is never lost. So even in those ‘whoops didn’t mean to do that" moments that all users can have (especially in a lab), we can simply flag the record back as 'undeleted'.

Only selected users in the FHCRC network are given access to the system, and should an employee become terminated or if they leave the organization, their database access stops the second their network security is revoked. If that is not fast enough, we can go in and manually remove access to the specific user in a timely manner.

Our centralized database provides full backups daily and then creates logs of inserts and deletes every hour between each back up. Even if the Vidro database was to fail, often a fresh log file can be salvaged. This is because the transaction files and backups reside on different physical drives than the database itself, meaning no data are lost. Even if that is not the case, the worst case scenario possible is the database would lose up to 59 minutes worth of data. The chances of this disaster are rare, because the drives use RAID, which creates multiple mirrors of the data on the primary drives. Meaning if one drive were to fail, the mirror can immediately be used to take over. If the server were to fail (the machine blows its mother board or other internal part), the data are still intact as the drives are physically separated from the processors that work with it, and communicate to the processor via a fiber-optic channel.

We recently upgraded our primary production database server to increase its disk capacity to 750 Gigabytes.

Metrics:

Vidro currently tracks 119,351 individual samples and can match each sample to its master record and has typed each sample record into component, sub-component and sample group categories. All the DNA samples know which sample (be it buffy coat, buccal or tissue) created that DNA and all DNA dispatches know which DNA sample the dispatch came from. We can manage dispatches of all component types.

There are currently over 1.1 million attributes related to our samples. The total number of unique data points in Vidro is well over 2 million.

Despite this volume, search results of sample records, integrated with genetic test results, participant and study variables, gender, risk, sample type and other factors take on average 2 seconds to complete and currently reports seven unique data views of that data, as well as the ability to report the findings to spreadsheets at the click of a button. The software requires a seven to eight second load time upon initial load of the search screen to accommodate those searches; thereafter, searches are most often completed in less than 2 seconds.

Diagrams:

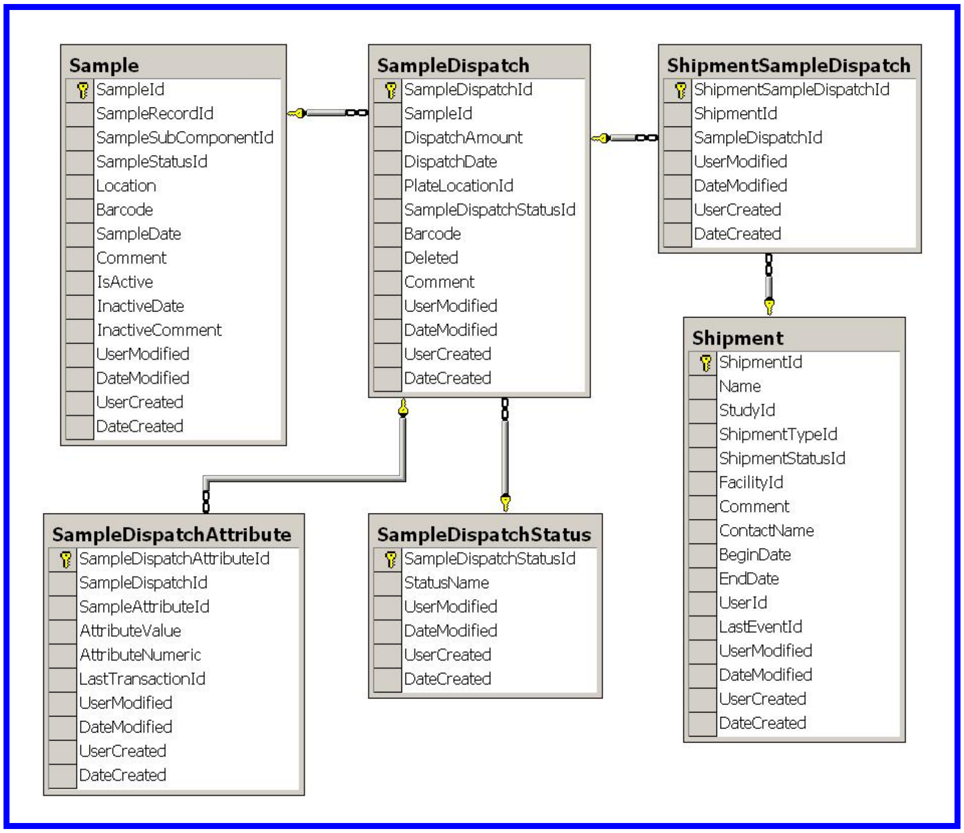


Diagram 1: Samples & Dispatch Schema

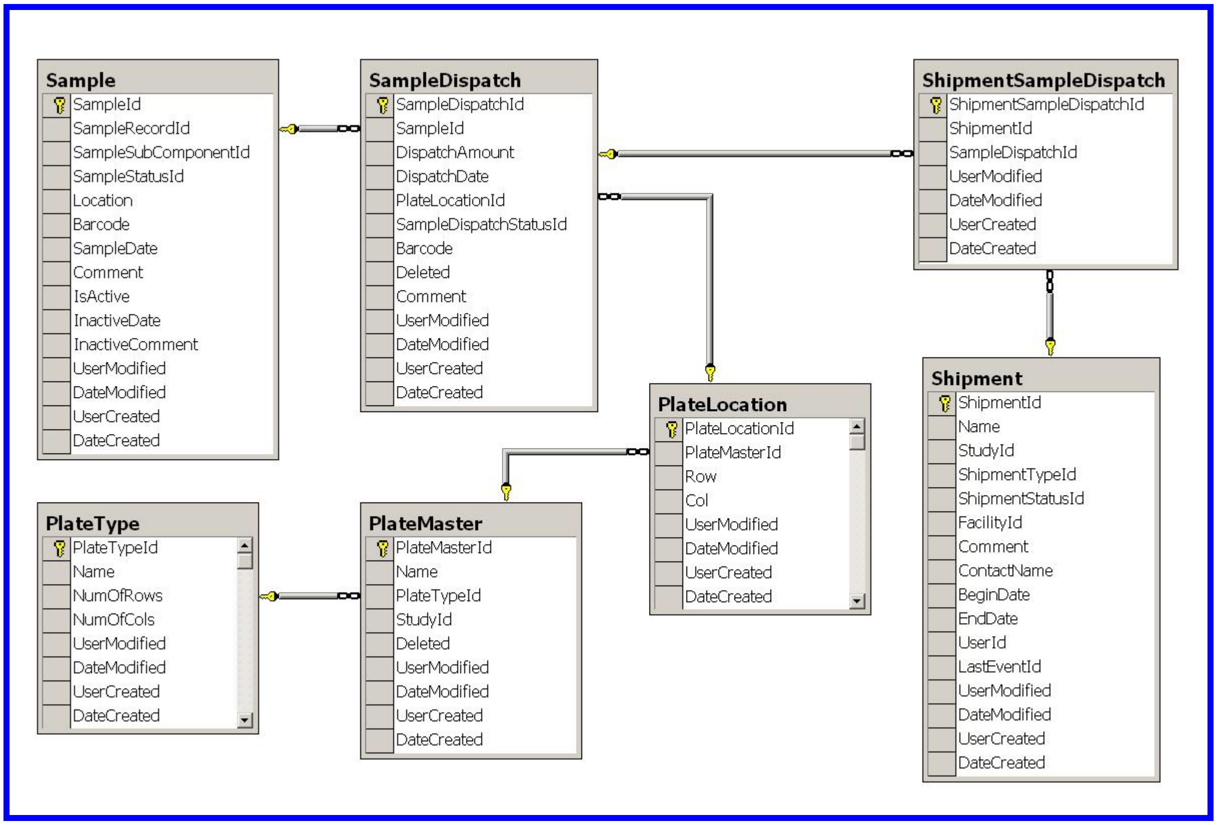


Diagram 2: Dispatch and Plating Schematics

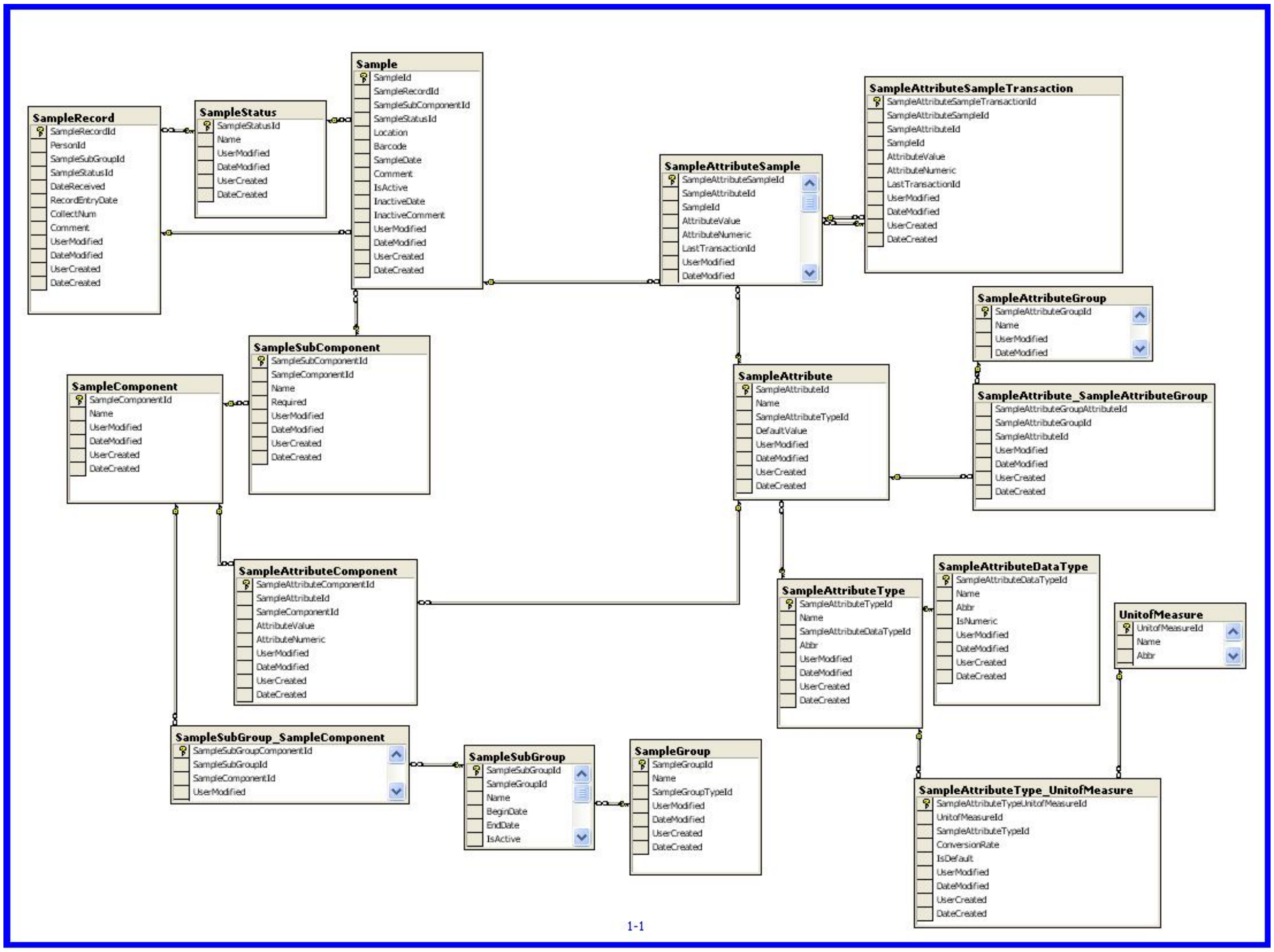


Diagram 3: Sample Classification and Attributes

USC C-CFR Biorepository Database
(Genetic Epidemiology Core)

Database Overview:

This new database was designed to accommodate the growing biorepository managed by Dr. Haile's lab. This system is designed using Microsoft SQL Server & Microsoft ASP.NET web development application. The advantages of this system are the robust design ability of SQL and the ease of the ASP.NET web-based user interface. This system can manage the collection of all biospecimens, catalog dispatches and maintain an ongoing inventory of each sample.

Database Features:

Blood: Collection, Processing & Dispatch

☑ A Blood Log that tracks and records daily delivery of blood

☑ Record the arrival time of blood

☑ Record the name of each technician working on blood processing

☑ Record the completion time of blood processing

☑ Record ‘Study_ID” submitted by collection center

- Synonymous to “Person_ID”

☑ Generate a unique ‘Specimen_ID’ for each blood received.

- This ID is applied to subsequent blood components: Blood Spots, Plasma, BC/RBCs, Lymphocytes and purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample generated is given a unique ‘Barcode_ID’

☑ Each sample is given a unique ‘Storage_Location’ that identifies its freezer, shelf, rack, box and well

☑ Record if sample is collected from ACD or EDTA tubes

☑ Record the condition of blood received, both ACD and EDTA tubes, and processing outcome.

☑ Record the reasons for pending redraw request

☑ Track the number of blood draws received from each patient.

☑ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples requested

- ○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample requested

☑ Records confirmation of dispatch receipt

☑ Records start date of dispatch

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request.

- ☑ DNA samples that are less than 2 ug/ tube will be flagged as "low inventory"
  - o These flagged aliquots will not be selected for any future dispatches

Paraffin Embedded Tissue: Collection, Processing and Dispatch

* Note that this branch of biospecimen collection is currently managed using Microsoft Access and Microsoft Excel. These data will be migrated to the same SQL database currently used for blood.

☑ A PET Log that records receipt of PET blocks and/or slides

☑ Record the arrival time of PET blocks and/or slides

☑ Record the completion time for block processing

☑ Record 'Study_ID" submitted by collection center

- ○ Synonymous to "Person_ID"

☑ Record 'Pathology_ID' given by hospital

☑ Record 'Block_ID' given by hospital

☑ Generate a unique 'Specimen_ID' for each PET block or batch of slides received.

- ○ This ID is applied to subsequent block components: Slides, paraffin-ribbons, TMA cores and purified DNA. This is the linkage variable throughout this portion of the database.
- ○ This ID can be matched to the hospital generated 'Block_ID'

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer or cabinet, shelf, rack, box, and well.

☑ A Dispatch Log that records biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples requested

  ○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample requested

☑ Records confirmation of dispatch receipt

☑ Records start date of dispatch

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory (number of slides, paraffin-ribbons, and TMA cores remaining) is updated based on dispatch request.

Australia C-CFR Biorepository Database

Database Overview:

This database was designed to accommodate the growing Australasian Colorectal Cancer Family Study (ACCFS) biorepository under the PI's John Hopper and Jeremy Jass. The biorepository and database are managed by Dr. Joanne Young's lab at QIMR. This system is designed using Microsoft Access. This system can manage the collection of all biospecimens, catalog dispatches and maintain an ongoing inventory of each sample.

Database Features:

Blood: Collection, Processing & Dispatch

☑ A Blood Log that tracks and records daily delivery of blood

☑ Record the collection date and arrival date of blood

  ☑ Record the name of each technician working on blood processing

  ☑ Record the Participant Number (NIH ID)and Date of Birth

  ☑ Sample given a unique Laboratory Number

☑ Record if sample is collected from ACD or EDTA tubes

  ☑ Information recorded on Access database

☑ Computer generates a unique 'Specimen_ID' for each blood received.

    ○ This ID is applied to subsequent blood components: Blood Spots, Plasma, Lymphocytes and purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer, shelf, rack, box and well

☑ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples dispatched

☑ Record number of samples dispatched

    ○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample dispatched

☑ Records confirmation of dispatch receipt

☑ Records completion date of dispatch (Shipment Date)

☑ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request and entered into the database.

Paraffin Embedded Tissue: Collection, Processing and Dispatch

*Note that this branch of biospecimen collection is currently managed using Microsoft Access and Microsoft Excel. These data will be migrated to the same SQL database currently used for blood.

☑ A PET Log that records receipt of PET blocks and/or slides and the date received and a unique Laboratory Number is given to the samples

☑ Record Participant ID (NIH_ID) submitted by collection center
  ○ Synonymous to "Person_ID"

☑ Record 'Block_ID' given by hospital and record the Pathology Laboratory from where the samples have been sent

☑ Generate a unique 'Specimen_ID' for each PET block or batch of slides received.
  ○ This ID is applied to subsequent block components: Slides, paraffin-ribbons, TMA cores and purified DNA. This is the linkage variable throughout this portion of the database.
  ○ This ID can be matched to the hospital generated 'Block_ID'

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer or cabinet, shelf, rack, box, and well. This includes the recording of Tissue stored as a back up resourse.

☑ A Dispatch Log that records biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples requested
  ○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample requested

☑ Records confirmation of dispatch receipt

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory (number of slides, paraffin-ribbons, and TMA cores remaining) is updated based on dispatch request on to the Access database

Ontario C-CFR Biorepository Database

Database Overview:

This database was designed to accommodate the growing biorepository at Mount Sinai Hospital. This system is designed using Microsoft SQL Server & Microsoft ASP.NET web development application. The advantages of this system are the robust design ability of SQL and the ease of the ASP.NET web-based user interface. This system can manage the collection of all biospecimens, catalog dispatches and maintain an ongoing inventory of each sample.

Database Features:

Blood: Collection, Processing & Dispatch

☑ A Blood Log that tracks and records daily delivery of blood

☑ Record the arrival time of blood

☑ Record the name of each technician working on blood processing

☑ Record the completion time of blood processing

☑ Record ‘Study_ID” submitted by collection center

- ○ Synonymous to “Person_ID”

☑ Generate a unique ‘Specimen_ID’ for each blood received.

- ○ This ID is applied to subsequent blood components: Blood Spots, Plasma, BC/RBCs, Lymphocytes and purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample generated is given a unique ‘Barcode_ID’

☑ Each sample is given a unique ‘Storage_Location’ that identifies its freezer, shelf, rack, box and well

☑ Record if sample is collected from ACD or EDTA tubes

☑ Record the condition of blood received, both ACD and EDTA tubes, and processing outcome.

□ Record the reasons for pending redraw request WE use a comment field.

☑ Track the number of blood draws received from each patient.

☑ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples **dispatched.**

☑ Record number of samples **dispatched.**

- ○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample **dispatched.**.

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request.

- ○ DNA samples that are less than 2 ug/ tube will be flagged as "low inventory"
- o These flagged aliquots will not be selected for any future dispatches

Paraffin Embedded Tissue: Collection, Processing and Dispatch

☑ A PET Log that records receipt of PET blocks and/or slides

☑ Record the arrival time of PET blocks and/or slides

☑ Record the completion time for block processing

☑ Record 'Study_ID" submitted by collection center

- ○ Synonymous to "Person_ID"

☑ Record 'Pathology_ID' given by hospital

☑ Record 'Block_ID' given by hospital

☑ Generate a unique 'Specimen_ID' for each PET block or batch of slides received.

- ○ This ID is applied to subsequent block components: Slides, paraffin-ribbons, TMA cores and purified DNA. This is the linkage variable throughout this portion of the database.
- ○ This ID can be matched to the hospital generated 'Block_ID'

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer or cabinet, shelf, rack, box, and well.

☑ A Dispatch Log that records biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples requested

☑ Record number of samples **dispatched.**

- ○ Selected by Study_ID or Number of samples needed

☑ Record amount of slides **dispatched.**

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory (number of slides, paraffin-ribbons, and TMA cores remaining) is updated based on dispatch request.

Tissue (Fresh frozen): collection, processing and dispatch

A Tissue Log that tracks and records daily delivery of tissue

☑ Record the arrival day of tissue

☑ Record 'Pathology_ID' given by hospital

☑ Record surgery date

☑ Record 'Study_ID" submitted by collection center

- ○ Synonymous to "Person_ID"

☑ Generate a unique 'Specimen_ID' for each tissue sample received.

- ○ This ID is applied to subsequent tissue components: purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample generated is given a unique 'Barcode_ID' In progress

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer, shelf, rack, box and well

☑ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of sample **dispatched.**

☑ Record number of samples **dispatched.**

- ○ Selected by Study ID or Number of samples needed

☑ Record amount of sample **dispatched.**.

☑ Records completion date of dispatch

☑ Upon completion of dispatch biospecimen inventory amounts (milligrams) are updated based on dispatch request.

Lymphoblast cell line (LCL): Collection, Storage & Dispatch

☑ Record date blood product is sent to our Tissue Culture laboratory for cell transformation.

☑ A LCL Log that tracks and records receipt of cell lines from our Tissue Culture laboratory.

☑ Record 'Study ID" submitted by collection center.

- ○ Synonymous to "Person_ID"

☑ Generate a unique 'Specimen_ID' for each LCL received.

- ○ This ID is applied to subsequent blood components: cryopreserved Lymphoblasts and purified DNA. This is the linkage variable throughout this portion of the database.

☑ Each sample is given a unique 'Storage_Location' that identifies its freezer, shelf, rack, box and well

☑ Record is transformation passed or failed.

☑ Record mycoplasma testing result for each cell line.

☑ Record outcome of freeze recovery of each vial of each LCL tested.

☑ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☑ Record destination of biospecimens for dispatch

☑ Record type of samples **dispatched.**

☑ Record number of samples **dispatched.**

○ Selected by Study_ID or Number of samples needed

☑ Record amount of sample **dispatched.**.

☑ Records completion date of dispatch

Appendix A.
Standard Biospecimen Tracking Checklist

Blood: Collection, Processing & Dispatch

☐ A Blood Log that tracks and records daily delivery of blood

☐ Record the arrival time of blood

☐ Record the name of each technician working on blood processing

☐ Record the completion time of blood processing

☐ Record 'Study_ID" submitted by collection center

  ○ Synonymous to "Person_ID"

☐ Generate a unique 'Specimen_ID' for each blood received.

  ○ This ID is applied to subsequent blood components: Blood Spots, Plasma, BC/RBCs, Lymphocytes and purified DNA. This is the linkage variable throughout this portion of the database.

☐ Each sample generated is given a unique 'Barcode_ID'

☐ Each sample is given a unique 'Storage_Location' that identifies its freezer, shelf, rack, box and well

☐ Record if sample is collected from ACD or EDTA tubes

☐ Record the condition of blood received, both ACD and EDTA tubes, and processing outcome.

☐ Record the reasons for pending redraw request WE use a comment field.

☐ Track the number of blood draws received from each patient.

☐ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☐ Record destination of biospecimens for dispatch

☐ Record type of samples **dispatched.**

☐ Record number of samples **dispatched.**

  ○ Selected by Study_ID or Number of samples needed

☐ Record amount of sample **dispatched.**.

☐ Records completion date of dispatch

- □ Upon completion of dispatch biospecimen inventory amounts (nano-gram, micro-liter) are updated based on dispatch request.
    - ○ DNA samples that are less than 2 ug/ tube will be flagged as “low inventory”
    - o These flagged aliquots will not be selected for any future dispatches
      Paraffin Embedded Tissue: Collection, Processing and Dispatch
- □ A PET Log that records receipt of PET blocks and/or slides
- □ Record the arrival time of PET blocks and/or slides
- □ Record the completion time for block processing
- □ Record ‘Study_ID” submitted by collection center
    - ○ Synonymous to “Person_ID”
- □ Record ‘Pathology_ID’ given by hospital
- □ Record ‘Block_ID’ given by hospital
- □ Generate a unique ‘Specimen_ID’ for each PET block or batch of slides received.
    - ○ This ID is applied to subsequent block components: Slides, paraffin-ribbons, TMA cores and purified DNA. This is the linkage variable throughout this portion of the database.
    - ○ This ID can be matched to the hospital generated ‘Block_ID’
- □ Each sample is given a unique ‘Storage_Location’ that identifies its freezer or cabinet, shelf, rack, box, and well.
- □ A Dispatch Log that records biospecimen request for both internal processing of samples and external sample request by approved collaborators
- □ Record destination of biospecimens for dispatch
- □ Record type of samples requested
- □ Record number of samples **dispatched.**
    - ○ Selected by Study_ID or Number of samples needed
- □ Record amount of slides **dispatched.**
- □ Records completion date of dispatch
- □ Upon completion of dispatch biospecimen inventory (number of slides, paraffin-ribbons, and TMA cores remaining) is updated based on dispatch request.

Tissue (Fresh frozen): collection, processing and dispatch

A Tissue Log that tracks and records daily delivery of tissue

☐ Record the arrival day of tissue

☐ Record ‘Pathology_ID’ given by hospital

☐ Record surgery date

☐ Record ‘Study_ID” submitted by collection center

  ○ Synonymous to “Person_ID”

☐ Generate a unique ‘Specimen_ID’ for each tissue sample received.

  ○ This ID is applied to subsequent tissue components: purified DNA. This is the linkage variable throughout this portion of the database.

☐ Each sample generated is given a unique ‘Barcode_ID’ In progress

☐ Each sample is given a unique ‘Storage_Location’ that identifies its freezer, shelf, rack, box and well

☐ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

☐ Record destination of biospecimens for dispatch

☐ Record type of sample **dispatched.**

☐ Record number of samples **dispatched.**

  ○ Selected by Study ID or Number of samples needed

☐ Record amount of sample **dispatched.**.

☐ Records completion date of dispatch

☐ Upon completion of dispatch biospecimen inventory amounts (milligrams) are updated based on dispatch request.

Lymphoblast cell line (LCL): Collection, Storage & Dispatch

☐ Record date blood product is sent to our Tissue Culture laboratory for cell transformation.

☐ A LCL Log that tracks and records receipt of cell lines from our Tissue Culture laboratory.

☐ Record ‘Study ID” submitted by collection center.

  ○ Synonymous to “Person_ID”

□ Generate a unique ‘Specimen_ID’ for each LCL received.

  ○ This ID is applied to subsequent blood components: cryopreserved Lymphoblasts and purified DNA. This is the linkage variable throughout this portion of the database.

□ Each sample is given a unique ‘Storage_Location’ that identifies its freezer, shelf, rack, box and well

□ Record is transformation passed or failed.

□ Record mycoplasma testing result for each cell line.

□ Record outcome of freeze recovery of each vial of each LCL tested.

□ A Dispatch Log that tracks biospecimen request for both internal processing of samples and external sample request by approved collaborators

□ Record destination of biospecimens for dispatch

□ Record type of samples **dispatched.**

□ Record number of samples **dispatched.**

○ Selected by Study_ID or Number of samples needed

□ Record amount of sample **dispatched.**

□ Records completion date of dispatch