1. PURPOSE

This document describes the processes by which data from Masonic Cancer Center (MCC) data management applications and external sources are integrated, stored, and disseminated by Oncology Medical Informatics & Services (OMIS). Refer to Appendix A for a representative list of data sources.

Requirements for data integration may originate from clinical trial-/study-specific data management plans, grants, registries, and other research efforts, etc.

Related policies and processes are outlined in companion documents and any of their associated agreements:

- **SOPs:**
  - Application Access SOP (1001.n) – describes the processes through which users receive authorization and access to use data management applications
  - Clinical Trial/Study Data Management Plan (DMP) SOP (702.n) – describes the elements contained in study-specific data management plans which function as implicit requests for data integration and release

- **Policies:**
  - MCC Oncology Business Layer (OBL) Source Data Integration Policy (1004.n) – states the policy under which data is integrated in the OBL and the agreement with the respective Data Owner/Data Steward to obtain such data
  - MCC Data Use and Sharing Policy (1003.n) – describes 1) process for a Data Owner/Steward to request release of or access to data by a third party, and 2) the agreement between the Data Owner/Steward and that third party surrounding the use of that data
Key concepts as used within this document: (additional definitions in Appendix C)

Data refers to output from sources for which OMIS has accepted data custodianship responsibility. These data sources are either owned by MCC or authority has been given by the Data Owner (or a designated Data Steward acting on behalf of the Data Owner) to OMIS, via a data request or other approval mechanism, to serve as the Data Custodian.

Data considerations include:
- Sources of data
- Types of data
- Data ownership

Integration refers to the act of compiling data elements from various source systems and redefining them within a more global context, essentially “freeing each” from its source application and better managing accessibility based on business requirements.

Integration considerations include:
- What data sources are being integrated?
- Why are the data sources being integrated?
- What rules and methods are being used to integrate the data sources?

Use refers to the utilization of data that has been integrated according to the processes outlined in this SOP.

Use considerations include:
- Who is going to use the integrated data?
- Governance and management of rights for access to and use of data; may involve creation of audit trails
- Data dissemination from the Oncology Business Layer (OBL) in the form of data sets and reports

Sharing refers to the act of making data available through an approved governance process to an authorized party.

Sharing considerations include:
- Identifying the Data Owners/Stewards who are sharing specified data
- With whom data will be shared and which data will be shared
- How sharing permissions will be managed to limit access to authorized parties
- Includes sharing data for mandated regulatory submissions
2. SCOPE

This SOP covers procedures necessary for OMIS to integrate data and perform custodial duties in relation to the use and sharing of this data:

- Adding or changing a data source
- Creating rules for data movement between architectural layers
- De-identifying data
- Designating source of truth
- Establishing data mapping
- Maintaining metadata
- Performing testing and quality assurance
- Conducting release management
- Managing storage and retention of data

This SOP does not cover procedures necessary for OMIS to incorporate data from OBL into data management applications. These types of needs will be addressed within data management application requirements.
3. RESPONSIBILITY

This section of the document specifies responsibilities with respect to data integration, storage, and dissemination processes and in accordance with Data Governance Roles and Responsibilities described in Appendix D.

3.1 OMIS will be the Data Custodian on behalf of the Data Owners and Data Stewards, and is responsible for ensuring compliance with the processes described in this SOP.

3.2 Management and/or Data Owners and Data Stewards, such as principal investigators, have responsibility for ensuring that all individuals who have access to PHI data have required HIPAA training and have legitimate work-related need for access to the data. They are also responsible for notifying OMIS when a person is no longer authorized to use an application.

3.3 Data Stewards and Data Owners are responsible for user acceptance testing and approval for release of data from the OBL. They are also responsible for resolving data quality issues identified through the data integration process.

3.4 All end users of reports and data generated from OMIS-managed applications will be responsible for the data they receive and may share it only with personnel who have a legitimate work-related need for the information. When custodianship moves from OMIS to end users, these users assume responsibility for the data’s security and dissemination in compliance with MCC or Academic Health Center (AHC) policy and all federal and state laws.

4. PROCEDURES

This portion of the document provides information about procedures that correspond with the topics outlined in the document’s scope section. Information is provided within this section for each topic and, where separate documents exist, they are also referenced.

4.1 Add a new data source

4.1.1 Complete a Data Source Survey (refer to “Governance: Data Source Survey” in Confluence) and update the Data Source Survey Grid which serves as a consolidated reference document for all data sources.

4.1.2 Develop an MCC OBL Data Integration Agreement between OMIS and Data Owner(s)/Steward(s) and obtain approval (refer to “MCC OBL Data Integration Agreement” in Confluence).

4.1.3 As needed, complete the Data Security Checklist based upon the “Draft NIST Security Architecture Design Process for Health Information
Exchanges” available in Confluence. Note that the caBIG Data Sharing and Security Framework (DSSF) offers additional resources for consideration.

When data sources become available from the Academic Health Center’s planned Trusted Independent Data Environment (AHC-TIDE), they will be treated as either new (4.1) or changed (4.2) sources as applicable.

4.2 Change an existing data source

Changing a data source may involve revising an existing data source, such as adding new data fields. It can also involve replacement of an existing data source with a new source.

When changing a data source, the Data Source Survey, Data Source Integration Agreement, and Data Security Checklist will be reviewed and revised as needed. Then all other procedures below will be followed as they apply to the particular change.

When a source of data, such as a software application, is no longer active, the data source integration process to acquire data from that data source will be retired. All data which had been acquired from that data source will continue to reside in the OBL (e.g., legacy tables).

4.3 Create rules for data movement between architectural layers

Rules for data movement are created in response to business and technical requirements for data.

4.3.1 Source to landing

Policies:
- Full refreshes and incremental loads as needed
- Landed data will be retained based upon regulatory requirements
- Standards will be established as needed: data object naming, etc.

4.3.2 Landing to transformation

Policies:
- Data will be loaded according to the frequency with which source data is landed and integrated data is needed in the OBL

4.3.3 Transformation to OBL – is dependent upon business rules and technical requirements for data integration. When needed to support data transformation processes, intermediate results will be stored in the transformation layer.
4.3.4 OBL to application – specifications are obtained from user groups, PI requests, particular grant requirements, and possibly other sources

4.3.5 Creation of data sets for biostatisticians – refer to Clinical Trial/Study DMP SOP or other document as appropriate

4.3.6 Frequency of data movement is dependent upon the rate of change of source data and end user requirements.

4.4 De-identify data

4.4.1 Create limited data sets

Limited data sets are created in accordance with the applicable University Administrative Procedure:

- Creating and Disclosing a Limited Data Set
  (http://www.policy.umn.edu/Policies/Operations/Health/HIPACOMPONENT_PROC10.html)

- Creating and Disclosing a Limited Data Set for Research
  (http://www.policy.umn.edu/Policies/Operations/Health/HIPARESEARCH_PROC03.html)

4.4.2 Create fully de-identified data sets

Fully de-identified data sets are created in accordance with the University Administrative Procedure De-identifying Data for Research
(http://www.policy.umn.edu/Policies/Operations/Health/HIPARESEARCH_PROC04.html)

4.5 Designate source of truth – in collaboration with Data Owners and Data Stewards, when synonymous data is collected by OMIS from multiple systems, an evaluation and determination is made of which system is considered to be source of truth; may involve defining precedence of sources

4.6 Establish data mapping

Data mappings essentially serve as a cross reference between separate data models. These mappings establish a semantic link, for example, between an MCC data element and the corresponding item in another system such as Epic. Data mappings will be stored in the metadata repository and may be made available in a data dictionary or OBL as needed.

4.6.1 Mapping to standards (e.g., caDSR, SNOMED CT) is performed as needed.

4.6.2 Mapping to external systems (e.g., CIBMTR Stem Cell Transplant Outcome Database) is performed as needed.
4.7 Maintain metadata

Two broad categories of metadata are maintained: business and technical. Maintenance of business and technical metadata involves creating metadata for new sources as they are acquired and modifying as needed.

4.7.1 Data element definitions will be curated with assistance of subject matter experts

4.7.2 Data dictionary entries may be entered directly into the metadata repository or transferred into the repository from application-specific entries collected in the Data Dictionary Template or obtained through reverse engineering of existing sources

4.7.3 Data element lineage to provide traceability of data from its original source(s) to its releasable form

4.7.4 Data mapping rules – refer to item 4.6

4.8 Perform testing and quality assurance

4.8.1 Monitoring will be performed at all layers of the data architecture to ensure that data is not refreshed when source data is unavailable or some type of failure has occurred.

4.8.2 QC metrics – includes data quality profiling performed in BI applications

4.8.3 Identify issues – identification through formal testing and through feedback from user groups and Data Stewards

4.8.4 Facilitate corrections when possible through Data Stewards and designate instances where data cannot be corrected, such as when an inconsistent identifier has been issued by an IRB

4.9 Conduct release management

This portion of the procedure includes change management and version control for architectural artifacts including load scripts, SSIS packages, and database schemas used for production applications and the transformation and OBL architectural layers.

4.9.1 Change management

4.9.2 Version control
4.10 Manage storage and retention of data

4.10.1 Daily snapshots are taken of data sources within the landing environment. These snapshots are transmitted to Autonomy, a company within Hewlett-Packard, using a secure web service. (This service was initiated with Iron Mountain which was subsequently acquired by HP in 2011.)

4.10.2 Retention schedule – refer to the “Research Data Retention” page in Confluence and the National Institutes of Health (NIH) Guidelines for the Conduct of Research.
5. EXHIBITS

The Exhibits section contains snapshots of documents, which exist elsewhere. Excerpts of those documents are included here for illustrative purposes only and are subject to change. Refer to the applicable source document for the most current information.

5.1 MCC Data Integration Flow Chart

<table>
<thead>
<tr>
<th>Data Integration Process 01-25-2012</th>
<th>Data Integration and Review</th>
<th>Data Dissemination</th>
</tr>
</thead>
<tbody>
<tr>
<td>Source Identification</td>
<td>Identify new data source</td>
<td>User acceptance agreements and (IRB) approvals (if needed)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Access controls, ETL data dictionary, business rules, quality control</td>
</tr>
<tr>
<td></td>
<td>Provide or validate data definitions</td>
<td>Review integrated data, QC</td>
</tr>
<tr>
<td></td>
<td>Prepare Data Source Integration Agreement</td>
<td>Data retention implemented</td>
</tr>
<tr>
<td></td>
<td>Data Owners / Data Stewards / SMEs</td>
<td>Dissemination, consume integrated data, prepare for publishing</td>
</tr>
<tr>
<td></td>
<td>OMIS / CRS Manager</td>
<td>Pl / Stats</td>
</tr>
</tbody>
</table>


5.2 Data Source Survey Template

**Governance: Data Source Survey**

*(These answers should be filled out as much as possible ahead of time with IT staff's current understanding of the systems, to be verified or corrected as needed by the Data Owner/Data Steward)*

1. Source system name:

2. Identify current roles
   a. Data Owners
   b. Data Stewards
   c. Data Custodians

3. Who currently has access to data?
   a. How are they authenticated?
   b. How are they authorized?
   c. What levels of authorization are used?

4. Describe current data flow process

Conflience: [https://confluence.cancer.umn.edu/download/attachments/8258026/DG_SourceSurvey.doc?version=3&modificationDate=1327002148000](https://confluence.cancer.umn.edu/download/attachments/8258026/DG_SourceSurvey.doc?version=3&modificationDate=1327002148000)

5.3 Data Source Survey Grid

<table>
<thead>
<tr>
<th>System</th>
<th>Data Owner(s)</th>
<th>Data Stewards</th>
<th>Data Custodians</th>
<th>Planning Groups</th>
<th>Others who access data</th>
</tr>
</thead>
<tbody>
<tr>
<td>BMT</td>
<td>BMT Program</td>
<td>Valerie O'Brien*</td>
<td>CMI staff</td>
<td>BMT Exec Grp</td>
<td>CRAs</td>
</tr>
<tr>
<td></td>
<td>Dan Weidendorf</td>
<td>CDA team</td>
<td>CTO staff</td>
<td>BMT Strategy Grp</td>
<td>Doctors</td>
</tr>
<tr>
<td></td>
<td>John Wagner</td>
<td>Todd DeFor†</td>
<td>CDAs</td>
<td>BMT User Grp</td>
<td>Nurses</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Qing Cao</td>
<td>Data Entry Staff</td>
<td></td>
<td>CTO Regulatory Staff</td>
</tr>
<tr>
<td>TIPS</td>
<td>Jeff Miller</td>
<td>Julie Curtissinger**</td>
<td>OMI Staff</td>
<td></td>
<td></td>
</tr>
<tr>
<td>LIMS</td>
<td></td>
<td>Rose Wangen</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Authorizes Access
† Responsible for Quality

Conflience: [https://confluence.cancer.umn.edu/download/attachments/8258026/DG_Source_SurveyGrid.xlsx?version=6&modificationDate=1330551239000](https://confluence.cancer.umn.edu/download/attachments/8258026/DG_Source_SurveyGrid.xlsx?version=6&modificationDate=1330551239000)
5.4 MCC OBL Data Integration Agreement Template

MCC OBL Data Integration Agreement
Between
Masonic Cancer Center Oncology Medical Informatics & Services
and
Data Owner/Steward(s): ____________

This agreement outlines specific roles and responsibilities shared by the University of Minnesota MCC OMIS and the Data Owner/Steward in accordance with the MCC OBL Source Data Integration policy.

The Data Owner/Steward and MCC OMIS staff may amend this Agreement by mutual consent, in writing, at any time.

Data Source name (or insert description):

Contained in MCC OBL Source Data Integration Policy

5.5 Data Security Checklist

This is the initial release, version 1.0, of the University of Minnesota Academic Health Center’s Information System Security Plan (ISSP). All controls from the seventeen security control families have been addressed at some level of detail. This is an evolving document and will be enhanced or modified as security controls are implemented and perfected.

The content of this document was compiled by:
- University of Minnesota, Academic Health Center – Information Systems
- University of Minnesota, Office of Information Technology – Security and Assurance
- University of Minnesota, Office of Privacy and Security

<table>
<thead>
<tr>
<th>CNTL NO.</th>
<th>Security Control Title</th>
<th>How Implemented or planned to be Implemented</th>
<th>Scoping Guidance</th>
<th>Common Control</th>
<th>Responsible Party</th>
</tr>
</thead>
<tbody>
<tr>
<td>AC-1</td>
<td>Access Control Policy and Procedures</td>
<td>AHC-IS maintains some policies that address access control for particular applications or Active Directory.</td>
<td>TBD</td>
<td>No</td>
<td>Academic Health Center Information Systems (AHC-IS)</td>
</tr>
<tr>
<td>AC-2</td>
<td>Account Management</td>
<td>Most accounts are currently maintained manually. Active Directory (AD) accounts are automated to be disabled or deleted. Audit trails for account creation are in place or planned to be added.</td>
<td>TBD</td>
<td>No</td>
<td>AHC-IS</td>
</tr>
<tr>
<td>AC-3</td>
<td>Access Enforcement</td>
<td>Applications and systems are designed to restrict privileged (administrative) functions by person or role.</td>
<td>TBD</td>
<td>Yes</td>
<td>AHC-IS</td>
</tr>
<tr>
<td>AC-4</td>
<td>Information Footprint and Enforceable Policies</td>
<td>Information flows controlled on the network via firewall at the</td>
<td>TBD</td>
<td>Yes</td>
<td>AHC-IS</td>
</tr>
</tbody>
</table>
5.6 Data Dictionary Template

<table>
<thead>
<tr>
<th>Name</th>
<th>Table</th>
<th>Data Element Name</th>
<th>Data Type</th>
<th>Definition</th>
<th>PHI Indicator</th>
<th>Permissible Values</th>
<th>Alias</th>
<th>Notes</th>
</tr>
</thead>
</table>

Confluence:
https://confluence.cancer.umn.edu/download/attachments/11141135/data+dictionary+template.xls
## Appendix A – MCC Data Sources

This list of sources will be modified as needed. This list consists of data sources as of this document’s most recent revision date. Data sources may be releasable to the OBL. A current list of MCC data sources is maintained in the Application Access and Data Governance Grid in Confluence.

<table>
<thead>
<tr>
<th>Source database</th>
<th>Primary usage</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>OnCore</td>
<td>Clinical trials regulatory data with eCRF and subject data, adverse event (AE) and serious adverse event (SAE) information</td>
<td>Oracle 11g</td>
</tr>
<tr>
<td>BMT</td>
<td>BMT outcome data for all transplant subjects</td>
<td>Oracle 11g</td>
</tr>
<tr>
<td>Fairview Cancer Registry (FCR)</td>
<td>Patient registry identified data (also known as NDB)</td>
<td>Oracle 11g</td>
</tr>
<tr>
<td>TTL LIMS</td>
<td>Analytes and subject test data</td>
<td>MS Access</td>
</tr>
<tr>
<td>TTL LIMS (archive)</td>
<td>Analytes and subject test data</td>
<td>MS Access</td>
</tr>
<tr>
<td>Miller LIMS</td>
<td>Analytes and subject test data</td>
<td>MS Access</td>
</tr>
<tr>
<td>Freezer Works</td>
<td>Sample inventory</td>
<td>4D</td>
</tr>
<tr>
<td>HL7</td>
<td>Lab data for on-study subjects</td>
<td>HL7 to SQL Server</td>
</tr>
<tr>
<td>HL7 - GU</td>
<td>Visit summary for all genitourinary (GU) patients; patients need not be on study</td>
<td>HL7 to SQL Server</td>
</tr>
<tr>
<td>TASCs</td>
<td>Finance data for nurse visits, lab charges, etc.</td>
<td>Oracle 11g</td>
</tr>
<tr>
<td>Monitoring</td>
<td>Subject visit and forms monitoring data</td>
<td>MS Access</td>
</tr>
<tr>
<td>BMT Odersets</td>
<td>BMT physician treatment orders</td>
<td>MS Access</td>
</tr>
<tr>
<td>Tobacco</td>
<td>Various databases with survey and bio-marker data</td>
<td>MS Access / Excel</td>
</tr>
<tr>
<td>Snus</td>
<td>Snus application for tobacco research</td>
<td>MS Access / Oracle</td>
</tr>
<tr>
<td>Pathology</td>
<td>Pathology files from Fairview</td>
<td>Text</td>
</tr>
<tr>
<td>Encounter (GU)</td>
<td>Encounter files for GU from UMP</td>
<td>Text</td>
</tr>
<tr>
<td>Encounter (NDB)</td>
<td>Encounter file for NDB from UMP</td>
<td>Text</td>
</tr>
<tr>
<td>MCT Product</td>
<td>Product data from MCT lab</td>
<td>MS Access</td>
</tr>
<tr>
<td>Source database</td>
<td>Primary usage</td>
<td>Type</td>
</tr>
<tr>
<td>---------------------------</td>
<td>---------------------------------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>P01 data</td>
<td>Jeff Miller’s P01, KIR and NMDP data</td>
<td>Text, Excel</td>
</tr>
<tr>
<td>LMTB</td>
<td>LMTB trial data</td>
<td>MS Access/Excel</td>
</tr>
<tr>
<td>UM Follow-up</td>
<td>Lance Armstrong Foundation</td>
<td>MS Access/Oracle</td>
</tr>
<tr>
<td>caTissue</td>
<td>P50 CENIC, with anticipated expansion to other uses</td>
<td>XML or csv</td>
</tr>
<tr>
<td>Steve Hecht lab</td>
<td>P50 CENIC</td>
<td>Excel</td>
</tr>
<tr>
<td>REDCap</td>
<td>P50 CENIC</td>
<td>.csv</td>
</tr>
<tr>
<td>Qualtrics</td>
<td>P50 CENIC</td>
<td>.csv</td>
</tr>
<tr>
<td>IVR</td>
<td>P50 CENIC</td>
<td>.csv</td>
</tr>
<tr>
<td>Johns Hopkins Cognitive Tasks software for CENIC</td>
<td>P50 CENIC</td>
<td>MySQL</td>
</tr>
<tr>
<td>CReSS</td>
<td>P50 CENIC puff topography</td>
<td>Text</td>
</tr>
</tbody>
</table>
Appendix B – Steps to Integration

This appendix contains an outline of typical data integration and dissemination steps and the sequence in which they are performed for a new data source. This level of detail may be of interest to Data Owners and Data Stewards who would like to obtain a high-level understanding of the overall data integration and dissemination process. This information is intended to complement the procedures specified in earlier sections of this document.

Step 1. Data Source Identification

- Name and define data source that needs to be integrated into the OBL or applications
- Gather high level requirements and data sharing and use agreement signed
- Confirm technical feasibility
- Confirm IRB approvals if necessary
- Confirm Consent if necessary
- Identify data stewards and owners through Data Source Survey

Step 2. Data Movement Process

- Define ETL (Extract Transform Load) rules
  - Integrate source to landing to create a mirror copy of data
  - Use rules to load OBL
  - Define frequency of data movement
  - Embed access controls
  - Embed QC
- De-identify if needed
- Create metadata, data dictionary
- Define source to landing data and landing to OBL data elements

Step 3. Data Review

- Engage data stewards/owners to review quality
- Make BI applications/reports as needed
- Embed access controls as needed
- Make baseline quality metrics
- Generate error reports and cycle back to source for correction
Corrections might result in process flow changes

Cycle back to quality metrics to show improvement in data

Step 4. Data Dissemination

- Make data available to users via BI applications/reports
- Generate data sets for analysis and publication
- Dissemination performed in accordance with MCC Data Use and Sharing Policy

Step 5. Data Retention

- Create a data retention plan
- Archive final data sets for 7 years (or as needed length of time)
### Term or Concept | Definition
--- | ---
architectural layers | A series of defined layers that constitute the OMIS data architecture, listed in the order of progression of data through the data architecture.

- **source** | The source layer consists of operational data sources associated with a production system or data feed, typically in the form of a relational database or file (e.g., csv, xls, xml). Data feeds may be recurring or one-time. Data sources are brought into the OMIS data architecture at a frequency consistent with the rates of change in their volume and structure.

- **landing** | The landing layer contains copies of production sources (including relational databases, flat files, and data in HL7 format). This layer replicates the schema of each source as closely as possible.

- **transformation** | The transformation layer contains intermediate forms of data as the data undergoes transformation from its native form in landing into a releasable form in the Oncology Business Layer.

- **Oncology Business Layer (OBL)** | The Oncology Business Layer consists of read-only data from various source systems that has progressed through the OMIS data architecture and is considered releasable by source system data stewards. The OBL includes normalized data and data which has been denormalized for the convenience of applications, reports, and other end uses of the data. OBL content is consumable by authorized individuals, applications, and tools.

- **business intelligence (BI)** | Business intelligence software is used to create front-end applications used to view data only. These specialized front-end applications, referred to as BI applications, are typically used for dashboards, reports, and analysis. BI applications retrieve data from one or more data sources.

- **data dictionary** | A data dictionary is a document containing information about a particular set of data such as its definition, format, origin, and usage.

- **data management application (DMA)** | MCC data management applications are front-end applications that connect to database management systems used for data entry to manage adding data or changing data entered into a DBMS. Data management applications are needed to support the functions performed by the University of Minnesota Cancer Center. Examples of data management applications include tracking protocols (active and inactive) and subjects enrolled in protocols, laboratory and tissue management applications, time tracking applications, tobacco research applications, and any database or application that MCC will use to conduct it business and mission.

- **data source** | A data source refers to any specified source of data: data within a management application, a database, a file, or other data collection. A data source may refer to a subset of data within an application, e.g., tobacco research data in caTissue application.

- **database management system (DBMS)** | A database management system is software whose primary purpose is to store and retrieve data, based on instructions from a front-end application or accessed directly using SQL queries.
<table>
<thead>
<tr>
<th>Term or Concept</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>de-identified data set</td>
<td>Per University policy, “Data is de-identified if, in accordance with the de-identification standard, 45 CFR §164.514(a)(b), a person with appropriate knowledge and skill uses accepted scientific principles and methods to determine the risk is very small that information could be used to identify the individual and documents the method used to justify such a conclusion; or if all of the following identifiers for the individual and the individual's relatives, employers or household members are removed” as described at <a href="http://www.policy.umn.edu/Policies/Operations/Health/HIPAARESEARCH_PROC04.html">http://www.policy.umn.edu/Policies/Operations/Health/HIPAARESEARCH_PROC04.html</a></td>
</tr>
<tr>
<td>front-end application</td>
<td>A <strong>front-end application</strong> is the user interface or that part of software or a website that a user sees on the screen and acts on for the purpose of entering commands or accessing other parts of the software or website. The front-end application may use a web browser or desktop user interface to access application data. This is in contrast to a back-end application, which manages direct access to database tables, such as a DBMS.</td>
</tr>
<tr>
<td>limited data set</td>
<td>A <strong>limited data set</strong> must meet the requirements stipulated in the applicable University Administrative Procedure:</td>
</tr>
<tr>
<td></td>
<td>• Creating and Disclosing a Limited Data Set (<a href="http://www.policy.umn.edu/Policies/Operations/Health/HIPAACPONENT_PROC10.html">http://www.policy.umn.edu/Policies/Operations/Health/HIPAACPONENT_PROC10.html</a>)</td>
</tr>
<tr>
<td></td>
<td>• Creating and Disclosing a Limited Data Set for Research (<a href="http://www.policy.umn.edu/Policies/Operations/Health/HIPAARESEARCH_PROC03.html">http://www.policy.umn.edu/Policies/Operations/Health/HIPAARESEARCH_PROC03.html</a>)</td>
</tr>
<tr>
<td>metadata</td>
<td><strong>Metadata</strong> is often referred to as “data about data”.</td>
</tr>
<tr>
<td></td>
<td>OMIS maintains two broad categories of metadata: business and technical.</td>
</tr>
<tr>
<td></td>
<td>• Business metadata consists of items such as data dictionaries and data element definitions. This metadata is intended to support all levels of users in understanding the data within the OBL and the applications or reports which receive their data from the OBL.</td>
</tr>
<tr>
<td></td>
<td>• Technical metadata consists of items such as security groups and server usage. This metadata is intended for use by OMIS personnel with responsibility for maintaining the technical architecture that comprises the data integration environment.</td>
</tr>
<tr>
<td>source of truth</td>
<td>When synonymous data is available from multiple sources, the location which is deemed by the Data Owner in consultation with OMIS to be the most authoritative is designated as the <strong>source of truth</strong>.</td>
</tr>
</tbody>
</table>
Appendix D – Data Governance Roles and Responsibilities

There are three essential roles in data management: Data Owner, Data Steward, and Data Custodian. It should be noted that these are roles, not job titles. Each role defines specific functions typically performed by different personnel in an organization. In some cases the same person can play multiple roles. While these roles are not mutually exclusive, they are hierarchical. A data owner may play the role of data steward or data custodian even if they are not explicitly defined in that role but the reverse is never true. A Data Custodian does not have the rights of a Data Owner or Data Steward except when explicitly designated. Just as a Data Steward may play the role of Data Custodian without explicit designation but not Data Owner. The only exception would be if they do not have the knowledge or skill set to competently accomplish the task associated with the role.

Data Owner
A Data Owner is a member of senior management that is ultimately responsible for the protection and use of the data. The Data Owner is responsible for the policy and practice decisions regarding the data, for its accuracy, and authorization for its access by others. An example of a Data Owner is the principal investigator of a study. A Data Owner typically assigns someone who has detailed knowledge of the subject matter and data requirements to support it to perform in the role of Data Steward.

Data Steward
The Data Steward is a subject matter expert who has authority to carry out activities to protect, establish policy and approve access to the data on behalf of the Data Owner. The primary focus for this role is managing data content and business logic.

Data Stewards may create or approve new policies and standards for data entry, usage, and transformation. They are responsible for quality control activities, such as defining metrics, data validation rules and reconciliation processes. Data Stewards work closely with Data Custodians to establish, maintain, and documents processes and standards, and monitors for compliance. When data quality errors are identified, Data Stewards are responsible for remediation including resolving the issue and determining new processes and ongoing monitoring.

Data Stewards may work within a customer group or within the OMIS team. Examples of Data Stewards include lab managers, research associates, postdoctoral fellows, statisticians, business analysts, etc.

Some Data Stewards are primarily concerned with defining the rules for point of data entry processes, capture and maintenance of business metadata, such as vocabulary definitions, valid values and business rules. Data Stewards also help define data access rights, validate business rules, data transformations and approve provisioning requests for access to source data related to the Oncology Business Layer (OBL).

Data Stewards may manage data access and reporting. In this role Data Stewards authorize and assist users to access the appropriate data from OBL.
Data Custodian

Data Custodians are primarily responsible for implementing the policies, processes and definitions as determined by Data Owners and Data Stewards. This includes data entry, and overseeing the safe transport, storage, and management of data. While content is important to them, their focus is on designing, deploying, and maintaining the underlying infrastructure and activities required to keep the data intact and available to users.

Responsibilities include assisting Data Stewards and other Data Custodians in researching data issues and inquiries, implementing data transformations, resolving data quality issues, and collaborating on system changes. They document all data movement, conduct data validation, and execute reconciliation processes following completion of data movement as specified by the Data Stewards.

At the data source level, Data Custodians provide source data, or access to source data within the OBL, provide the OMIS team and Data Stewards with appropriate communication of issues regarding delivery or quality of source data, and participate in the OBL planning process.

Data Custodians at the data repository level provide the data source custodians with appropriate communication of issues with delivery or quality of source data. They define, capture and maintain technical metadata.

Data Custodians also develop applications and reports to deliver data to authorized users according to the policies and business rules established by Data Owners and Data Custodians.

Examples of Data Custodians include programmers, DBAs, system administrators, report developers, etc.
Business Definitions of Data Governance Roles and Responsibilities

These definitions are intended for use when communicating with non-IT personnel.

There are three essential roles in data management: Data Owner, Data Steward, and Data Custodian. It should be noted that these are roles, not job titles. In some cases the same person can play multiple roles. A Data Owner may play the role of data steward or data custodian. For example, a PI (Data Owner) may collect data (Data Steward) in clinic and record it in Epic, then at a later time perform quality assurance review of reports involving this data (Data Steward).

Data Owner
Data Owners are responsible for the overall management of clinical trials and data management.

Data Steward
The Data Steward is a subject matter expert acting on behalf of the Data Owner responsible for ensuring the content and accuracy of the data.

Data Custodian
Data Custodians are primarily responsible for entering or collecting data. This includes data entry personnel, CRAs, and anyone else who performs data entry. Data Custodians are also responsible for implementation of case report forms (CRFs) for use in data collection.
SOP Number: 1002.02 - OMIS Infrastructure
Superseded SOP #: none
Functional Area: MCC Data Integration, Storage, and Dissemination
Applicable to: All data residing on MCC servers
SOP Title: MCC Data Integration, Use, and Sharing

SIGNATURES

Sandeep Kataria
Manager, Oncology Medical Informatics & Services
Masonic Cancer Center
University of Minnesota

Dr. Sarah Cooley, MD
Assistant Professor of Medicine
Director, Oncology Medical Informatics & Services
Associate Director, Cancer Experimental Therapeutics Initiative
Masonic Cancer Center
University of Minnesota

Effective Date: 06-12-2013
Date: 6/12/13

Date: 6/12/2013