Copy of caDSR - Overview of Use and Collaborations

This overview of use of and collaborations for the Cancer Data Standards Registry and Repository (caDSR) was initially prepared October 8, 2010 and is updated on an ongoing basis. The following topics are included:

- NCI/caBIG® Informatics Infrastructure - caDSR
  - NCI caDSR CDE Browser
    - File Downloads
  - NCI caDSR Form Builder
  - NCI caDSR Curation Tool
  - NCI cgMDR Excel Addin, Template and Bulk Loader
  - NCI caDSR UML Model Browser
  - NCI caDSR Sentinel Tool
  - NCI caDSR Database Server, Domain Model and Freestyle APIs
  - NCI caDSR Semantic Integration Workbench (SIW) and UML Loader
- NCI caCORE Training
- caDSR Usage in NCI and caBIG®
  - National Cancer Institute Cooperative Groups
  - Cancer Centralized Clinical Database (C3D)
  - NCI Cancer Therapy Evaluation Program (CTEP)
  - NCI Division of Cancer Prevention (DCP)*
  - NCI Center for Cancer Research (CCR)
  - Specialized Programs of Research Excellence (SPOREs)
  - The National Marrow Donor Program® (NMDP) and our Be The Match FoundationSM Project
calIntegrator
calB2B
  - The Cancer Genome Atlas (TCGA)
  - Imaging (In Vivo Imaging Workgroup, DICOM, ACRIN)
  - OpenMDR – Ohio State University
  - DECAMAP - Data Element Curation & Mapping Platform
evMerge - Data Element Curation & Mapping Platform
- caDSR Usage in NIH
  - National Heart Lung and Blood Institute (NHLBI)
  - National Institute of Dental and Craniofacial Research (NIDCR)
  - National Institute of Child Health and Development (NICHD)
  - National Institute of Neurological Diseases and Stroke (NINDS)
- caDSR External Adoption
  - Biopathology Center, National’s Children’s Hospital
  - PhenX Toolkit
  - GIATE
  - American Heart Association (AHA) and American College of Cardiology Foundation (ACCF)
  - Winthrop P. Rockefeller Cancer Institute at the University of Arkansas for Medical Sciences
  - Union of Light Ion Centers in Europe (ULICE)
  - Winthrop P. Rockefeller Cancer Institute, Cancer Control Core at the University of Arkansas for Medical Sciences; Breast Mammography
  - Oncology Patient Enrolment Network (OPEN) - Sponsor-CTEP/NCI
  - Duke Translational Medicine Institute
  - Taiwan Cancer Registry
  - Children’s Oncology Group (COG) and National Childhood Cancers Foundation
  - University of Miami Ontology Modeling
  - Center for International Blood and Marrow Transplant Research (CIBMTR)
  - The Human Studies Database (HSDB)
  - Towards Unambiguous formal descriptions of cancer therapy experiments
  - Ontology based queries for caGrid Infrastructure
  - Information Management Services, Inc. (IMS) BIS, NAACCR and DCCPS/caSEER program
  - The CTSA Health Ontology Mapper
  - A Growable Network Information System (AGNIS)
cancerGrid, cgMDR UK
  - CDC, PHIN Reportable Conditions and Common Core Data Elements
- Related Software Engineering Research Highlights
  - "Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network"
- Appendix A - Statistics
  - Data Element Descriptions
  - Case Report Forms and Surveys
  - UML Models
  - caDSR Tools and Browser
    - CDE Browser (for the month of June 2011)
    - Curation Tool (as of Feb 2011)
    - Form Builder (as of Feb 2011)
    - UML Model Browser (as of Feb 2011)
    - caDSR API services (as of June 2011)
    - Freestyle services (as of Feb 2011)
- Appendix B - Support for Clinical Trials
- Appendix C - Models
  - caBIG® (NCI cancer Biomedical Informatics Grid Data Services)
  - PS&CC (NCI Population Sciences & Cancer Control)
  - caCORE (NCI Core Infrastructure)
- Appendix D - Publications
nci Cancer Data Standards Registry and Repository (caDSR) is a centralized resource for services and web-based information technology tools to meet NCI needs for documenting and sharing unambiguous descriptions of data. These resources are used to develop standards-based IT descriptions for trials and research data shared by NCI and the larger international biomedical community, so data can be effectively and programmatically exchanged, aggregated, analyzed, used and reused for secondary research.

Since 2000, caDSR tools and content have become the primary means by which cancer researchers developing databases and clinical trials data managers describe the data collected in clinical and research studies:

- By NCI and its partners in caBIG®, NIH, other federal agencies, and U.S. and international biomedical and organizations, academia and cancer researchers, such as for documenting and publishing human and computer interpretable representation of standard datasets for cancer registries (NAACCR and SEER) and those used in clinical trials (Cancer Centralized Clinical Database (C3D));
- To document, harmonize access, and publish data descriptions bound to standard value sets biomedical terminologies published by EVS, such as BRIDG 3.0.2;
- For basic, translational, and clinical research, clinical care, epidemiology, public health, administration, and public information.

caDSR data descriptions, software and services are freely available where possible, promoting broader adoption but hindering full tracking of use; restrictions on identifying and surveying users have a similar effect. Like EVS, the caDSR adopters and patterns of use are fairly clear.

NCI/caBIG® Informatics Infrastructure - caDSR

One measure of caDSR usage is the number of data elements recorded by different communities for the purpose of recording the detailed descriptions of data, and enabling sharing of these descriptions among data managers.

caDSR provides the foundational layer for representing the precise meaning and representation of data, and through linkage to EVS terminology and ISO /IEC 11179 standard, the ability to translate the meaning throughout the informatics infrastructure and across various research domains.

Up until 2010, all data, (30k+ “Released” data elements, 44k+ total), and information models (146 including multiple versions of some models) for all systems and applications in caBIG® had to have their meaning coded with controlled terminology, and recorded in caDSR, a repository used by various stakeholder communities. The high number of data element detailed descriptions as caDSR metadata represent the many and varied ways in which similar data are collected and stored in databases throughout the community.

By exposing these details in a common, structured way, caDSR enables automated harmonization when constructing data collection instruments and software heretofore not possible or feasible because the details were hidden away in application programs and paper documentation. In addition to the recording the specific data elements used in applications, the registry also records the reusable semantic components that, when reused, provide the basis for detecting semantic equivalence and potential for data aggregation and common processing. There are 12,383 classes of information defined in caDSR that are reused to create over 30k Common Data Elements (CDEs). For example, the property “Email Address” is reused by 43 or more CDEs to describe data such as “Organization Email Address”, “Patient Email Address”, “Clinical Trial Participant Email Address”, “Investigator Email Address”, etc. This fact is recorded in the caDSR so that developers can detect this similarity between disparate data and enable them to discover and potentially combine data from different sources.

Other applications, such as caIntegrator and caB2B, access caDSR via API at runtime to display descriptive information about data on data collection forms or to provide drop down lists for populating fields. Others import CDEs to use to customize data collection forms in local software applications such as a Clinical Data Management System (CDMS), C3D, LimeSurvey and others.

Also worth noting, an open source XML-based version of caDSR was developed by the University of Oxford Computing Lab to support the cancerGrid's clinical trails and other clinical and population studies in the UK. caDSR CDEs were extracted and formed the basis of new content development, reusing many of the NCI's CDEs in UK trials.

Lastly, this XML-based registry was taken by Ohio State University and enhanced to work with caGrid, and provides the basis for annotating services with caDSR usage can be measured by the number of communities and applications that have registered their data elements and models in caDSR and are retrieving them via browser downloads and use of caDSR APIs. The caDSR software was designed by Oracle and the Census Bureau in the late 1990's as a centralized repository for ISO/IEC 11179 content, and due to its dependency on an Oracle database and an application server for the Admin Tool and database, is not well suited to installing small cancer centers who are resource-constrained or lack the Oracle skills to get the suite of products installed, customized and working. Instead, the NCI hosts content for those who want to use the infrastructure. Others who want their own registry now also have the cgMDR/openMDR option.

NCI caDSR CDE Browser

The CDE Browser is the starting point for exploring the details of the data elements described and registered in the caDSR. This tool supports browsing, searching, and exporting CDEs in XML or Excel format, within or across end user contexts as follows.

- Search for data elements by NCI Thesaurus Concept, Value Domain Permissible Value, Classifications or simple text searches in the names and definitions of the elements in caDSR.
- Using the Shopping Cart, create a customized set of CDEs for exporting or sharing with other applications in XML or Excel format.
- The CDE Browser publishes its DTD to export/download data elements in XML.
- Access the Tool at [https://cdebrowser.nci.nih.gov/CDEBrowser/](https://cdebrowser.nci.nih.gov/CDEBrowser/)
**Usage:** An average of 35 unique users visit the CDE Browser each day, with over 18,204 unique visitors since we started tracking statistics in 2005. The number of pages served during this time is 1,510,814. Approximately 11 pages were viewed per 133,149 visits. CTEP currently has over 10,000 released and draft CDEs. See Statistical Appendix for more details.

The NCI Cancer Clinical Research (CCR) center has 3,491 selected CDEs that are downloaded and imported into the Cancer Centralized Clinical Database (C3D), with 935,573 instances of CRFs built from these common descriptors and over 150M data points (responses) collected.

**Chart showing usage of CDE Browser**

![Chart showing usage of CDE Browser](chart.png)

**File Downloads**

Several collections of CDEs have been downloaded in Excel and EXML format from the CDE Browser and pre-packaged for downloading from NCI's Wiki. Pre-packaged downloads include all the CDEs that are in a "released" workflow status, those that are caBIG Standards and CDEs in NACCAR, SEER and BRIDG. Collections are posted to the caDSR Wiki on the [caDSR Hosted Data Standards, Downloads, and Transformation Utilities](https://wiki.cancer.gov/CASTLHD/caDSR/downloads) page.

**NCI caDSR Form Builder**

Form Builder organizes CDEs in forms that replicate the content of Case Report Forms. These forms can include complex form behavior, such as skip patterns. Using the [CDE Browser](https://formbuilder.nci.nih.gov), you can search for CDEs, place them in a shopping cart, and from there insert CDEs as Questions on a Form. As you place the CDEs on the Form, the tool uses the stored metadata to provide default question text and value domain information automatically. If the value domain information is presented as an enumerated list of values, you can perform basic functions to organize the list.

You can place one or more CDEs into one or more sections on the Form, organize them into groups, and save them as Modules that can be copied from one Form to another.

Key capabilities of Form Builder include:

- Define skip patterns between questions based on question responses
- Define repeating groups
- Define default values for questions in repeating groups
- Publish a Form in the caBIG® Context’s Form Catalog
- Subscribe to Sentinel Reports that are triggered by changes to CDEs on the Form
- Classify the Form in one or more caDSR Classification Schemes
- Download the Form to Excel
- View and print from a Printer Friendly version of the Form
- Edit, Save or Download the CDE Shopping Cart
- Forms shopping cart for storing collections of form to export to other systems


**Usage:** NCI has been tracking Form Builder usage since January of 2009, though the tool has been available since 2002. On an average day there are 6 curators using Form Builder, with 913 unique visitors over the past year (October 2009 to October 2010), viewing 14,546 pages. There are 407 Protocols defined in caDSR, into which many of the 2,613 CTEP Case Report Forms (CRFs) are grouped, and an additional 264 templates that describe minimum datasets to be collected for various types of cancer organized by disease and type of trial. NCI CTEP uses CDEs and CRFs for reporting trial results for CTEP Sponsored trials, of which there are currently 611 trials.

Commercial vendors Medidata (RAVE), Westat (for CTEP) and Eastern Oncology Centers Group (ECOG) retrieve CRFs using the caDSR API and import them to customize their data collection systems. Refer to the section titled NCI caDSR Database Server/Domain Model and Freestyle APIs for API usage.

ACRIN is involved in registering standard data elements for imaging and using them to create forms in Form Builder.
NCI caDSR Curation Tool

The CDE Curation Tool supports creation and editing of all the primary semantic descriptions for Data Elements used by the community in data repositories or application software. It is intended for use by Context Administrators. This tool's features facilitate the use of the CBIIT Enterprise Vocabulary Services (EVS) to create administered item names and definitions, helping to ensure ISO/IEC 11179 compliance and also use of caDSR naming conventions.

Public users can browse for CDEs, reusable ISO/IEC 11179 value domains (VD), data element concepts (DEC), Object Classes and Properties using identifiers, search strings, Classification Schemes or EVS concepts. The tool leverages the ISO 11179 metamodel to "get associated" items.


**USAGE:** Since 2005, on an average day 22 unique users use the Curation Tool, with 4,558 unique visits to the site, and over 1.6M pages served. New CDEs can be created using this tool, the Admin Tool or the UML Model Loader. With the detail of caBIG® models no longer required to be registered in caDSR, that combined with curator harmonization activities designed to ensure and increase reuse of existing CDEs, the number of new CDEs year-over-year is trending down.

NCI cgMDR Excel Addin, Template and Bulk Loader

The term “cgMDR” actually refers to a group of components designed to work together to help users get large lists of administered components bulk loaded into caDSR. The initial audience for these products was the National Marrow Donor Program (NMDP) but interest in its use has spread beyond that group.
cgMDR actually stands for the “CancerGrid Metadata Registry” and specifically refers to a downloadable localized database based on ISO/IEC 11179 where you can store and administer a personalized set of data elements and their components. This database and its interface were created by the CancerGrid team at Oxford University Computing Lab in the United Kingdom. The NCI GForge project archive can be found on this wiki.

The CBIIT cgMDR installation also includes a group of components, including two add-ins and two Excel 2007 spreadsheets that work in conjunction with cgMDR as well as with other data repositories. These additional features help provide a complete solution to assist you in creating a list of personalized data elements that can then be bulk loaded into caDSR without the need for creating a UML Model or by individual manual curation of each element.

**USAGE:** National Marrow Donor Program (NMDP), University of Michigan, ACRIN for batch loading CDEs and Value Domains into caDSR

### NCI caDSR UML Model Browser

The UML Model Browser supports web browsing and searching data described by UML Models transformed and loaded in the caDSR repository via the UML Loader. This allows users to find administered items that are part of registered UML models for data services on caGrid. The UML Model Browser supports browsing, searching and exporting the classes, attributes and relationships between classes of a UML domain model. Within the framework of a UML Model CDEs are mapped to the UML attribute level. Search results display the Package Names, Attributes and Java primitive types. The CDEs used for semantic resolution are presented as links to the CDE Browser. CDEs in UML Models can be downloaded from the UML Model Browser.

Access the Tool at [http://umlmodelbrowser.nci.nih.gov](http://umlmodelbrowser.nci.nih.gov)

**USAGE:** Since tracking began in March 2006, on an average day six visitors come to the site with 2,765 unique visitors through October 2010. 34,522 web pages were viewed in this four year period. There are 198 models loaded into caDSR describing the data for each of these applications using standard ISO /IEC 11179 descriptions. The descriptions of this data in caDSR ensure that the meaning of the data is unambiguously represented for both human and computer interpretation. The UML Model API is used by caGrid and caB2B to explore registered models programmatically. The caGrid Portal exposes the semantic metadata in its portal, which was accessed by 5,163 unique visitors between October 2009 and October 2010. According to ISO 8000, a Data Quality standard, registration of the details of these data and application models ensures the model owner both owns the data and meets guidelines for Data Quality.

### NCI caDSR Sentinel Tool

The caDSR Sentinel Tool was first introduced in 2005 to allow users to create and manage Alert Definitions for the caDSR. Alert Definitions are a set of rules that are periodically evaluated against the caDSR. If the conditions in those rules are met, notification is sent to the user by email, with a hyperlink to a report that specifies the changes that have taken place. A script that kicks off this tool runs nightly, but the reports can also be run through the user interface.

The caDSR Sentinel Tool provides the capability to:

- Monitor all changes to Administered Items including Data Elements, Data Element Concepts and Value Domains and Case Report Forms
- Filter report content by Context, Specific forms or templates, Classification Scheme, Class Scheme Item, Creator and Modifier
- Trigger report generation using Workflow Status, Registration Status and Version
- Set reports to automatically be generated daily, weekly, monthly or on demand
- Create a report distribution list which may optionally include a process URL to send the report in XML format to software for evaluation

Access the Tool at [https://cadsrsentinel.nci.nih.gov](https://cadsrsentinel.nci.nih.gov)

**USAGE:** On an average day, 3 users visit the caDSR Sentinel tool site with over 264 unique visitors between October 2009 and October 2010, and 8,354 page views in that time.

### NCI caDSR Database Server, Domain Model and Freestyle APIs

All caDSR content is available through various application programming and web service interfaces.

**caDSR API** allows users to access caDSR content by using a web browser to navigate the caDSR domain model and returns results in HTML or XML. A caDSR Java API provides a set of methods that can be used to retrieve content as XML documents. According to NCI statistics from Wusage, from October 2009 to October 2010, 628 unique visitors came to the site and accessed over 25M documents. The number of pages is slightly inflated due to the design of the caDSR Domain Model, as multiple pages between 6-20 pages must be served in order to get one logical document. Our estimate is that over 1.2M logical documents were retrieved.


**Freestyle API** is another caDSR interface that provides access to content both via a web browser and application programming interface. From October 2009 to October 2010, 328 unique visitors came to the Freestyle API site and accessed over 9.6k documents. The Freestyle API uses the caDSR Domain API to simplify object retrieval.

**Chart showing usage of caDSR API**

![caDSR API Pages in K](chart1)

**Chart showing usage of Freestyle API**

![Freestyle API Unique Visitors by Quarter](chart2)

**Usage:** The caDSR API usage information in the statistics appendix lists the top URLs that access the caDSR API. AGNIS uses the API to access Forms.
Illustration listing AGNIS information on using caDSR

From: 'Accessing AGNIS Metadata using the CaDSR'

NCI caDSR Semantic Integration Workbench (SIW) and UML Loader

The Semantic Integration Workbench (SIW) a tool that assists users in adding consistent metadata to a UML model represented as an XML file, or verifying consistency with existing caDSR content by tagging a domain model with matching concepts from the NCI Thesaurus, or attaching existing caDSR CDEs or Value Domains to attributes in the model. These annotations ensure reuse of CDEs or other metadata elements that have been previously recorded in caDSR. The UML Loader transforms the file and decomposes it into ISO/IEC 11179 descriptive metadata.

Access and use the Tool at [http://cadsrsiw.nci.nih.gov](http://cadsrsiw.nci.nih.gov) (Java WebStart application)

**USAGE:** As of July 2011, there are 7 new models loaded in 2011, 2 in the queue to be loaded and 2 on hold. 198 caBIG® and NCI UML models representing all versions of these applications' data have been transformed into ISO/IEC 11179 descriptive metadata in caDSR. These models use 26,360 CDEs. Many of the CDEs are reused across models. Thus, for example, if the CDEs were evenly distributed across all models, 133 CDEs would be used per model. However when searching for CDEs in each model, one finds that BRIDG v2.1 has 1,669 CDEs, caAERS v1.1 (Adverse Events Reporting System) has 463, caBIO 4.3 has 465, caTissue Suite 2.0 has 997, caNanoLab has 524 and C3PR v2.0 (Patient Registry) has 180. This includes several versions of BRIDG, caAERS and C3PR. See the Appendix for complete list of models with CDEs in caDSR.
The caDSR is part of an infrastructure that supports the semantic description of data. While it is based on ISO/IEC standards for internal representation of data elements, and makes it possible to interpret and compare meaning of data across different repositories and domains, it does not guarantee that the intended end users will immediately understand how to use this standard to describe their data, so training has been developed.

The [caCORE Training wiki](#) includes information about each course and self-paced workbooks, as well as a curriculum based on end user roles. This training was established in 2007 and has been utilized in training over 1400 caDSR end users. The primary courses are:

- Course 1000: Introduction to caCORE and caDSR
- Course 1010: Introduction to Metadata in the caDSR
- Course 1020: Using the CDE Browser
- Course 1025: Using the UML Model Browser (Created June 2010)
- Course 1030: Introduction to EVS
- Course 1040: Creating Well-formed Metadata and Metadata Business Rules
- Course 1045: Introduction to the Curation Tool (Created June 2010)

The CDE Browser is the primary public access for caDSR content. The course on using the CDE Browser is for advanced users to teach them how to leverage the advanced search and compare features. It has been completed by over 290 end users. According to web usage statistics, the CDE Browser site an average of 500 unique visitors each month.
USAGE: in addition to direct enrollments in caCORE and caDSR training, caDSR training materials are used in the AGNIS training materials.
caDSR Usage in NCI and caBIG®

National Cancer Institute Cooperative Groups

Project Sponsor: Dr. Jeff Abrams
Project Manager: —

The NCI Cancer Therapy Evaluation Program (CTEP) mandated that variables used in NCI-sponsored trials be registered in the caDSR several years ago. Using a series of unified workflow processes, all new and amended trials go through a review process that verifies the elements are added to the registry both as single elements, and in forms to represent a full clinical trial. There are more than 9600 data elements arranged programmatically in more than 2500 clinical trial-centric forms in the registry to represent the following Cooperative Groups:

- American College of Surgeons Oncology Group (ACOSOG)
- Cancer and Leukemia Group B (CALGB)
- Eastern Cooperative Oncology Group (ECOG)
- Gynecologic Oncology Group (GOG)
- European Organisation for Research and Treatment of Cancer (EORTC)
- National Cancer Institute of Canada (NCIC)
- National Surgical Adjuvant Breast and Bowel Project (NSABP)
- Radiation Therapy Oncology Group (RTOG)
- Southwest Oncology Group (SWOG)

An example of the content housed for each Cooperative Group is illustrated by ECOG content. ECOG is one of the largest clinical cancer research organizations in the United States, and conducts clinical trials in all types of adult cancers. Currently, ECOG has more than 90 active clinical trials in all types of adult malignancies. Annual accrual is 6,000 patients, with more than 20,000 patients in follow-up. These clinical trials use the caDSR to describe case report forms for sharing programmatically and human-understandable detailed descriptions of the data to be collected in clinical trials. This information is pulled electronically to use in customizing local software as well as publicly available for browsing or downloading descriptions via the caDSR CDE Browser.

Examples of ECOG clinical trials that have changed or improved cancer treatment methods include:

- A study that establishes an effective chemotherapy regimen in acute promyelocytic leukemia
- A study that demonstrated that a less-toxic regimen had similar results to the standard treatment of metastatic nonsmall cell lung cancer
A study that demonstrated that an intensive bone marrow transplant regimen for patients with metastatic breast cancer who responded to initial standard chemotherapy did not improve time to disease progression or lifetime survival.

A study that establishes an effective treatment program involving chemotherapy and radiation therapy for early lung cancer.

Cancer Centralized Clinical Database (C3D)

Project Sponsor: —
Project Manager: Christo Andonyadis (CBIIT), Dianne Reeves (CBIIT)

C3D is a clinical trials management system, for which Case Report Forms (CRFs) based on data elements (CDEs) are used to collect data. The CRFs and CDEs are described in caDSR and represented as computable metadata that is shared electronically to support the clinical center information systems required to collect data consistently and accurately across trials. This increases data accuracy, reduces clinician training and increases the possibility of aggregating data across studies.

C3D has been in operation since 2002, first in the NCI intramural research program (Center for Cancer Research (CCR)) and then adopted broadly across a growing number of groups. There are currently more than 315 trials from C3D represented in caDSR; reuse of content across trials and groups supports the need for interoperability, and ultimately the need to aggregate clinical trial data across studies and groups for safety and product reporting. In addition to CCR, C3D provides CDMS support for over 20 cancer centers (see Appendix B), for which trial CRFs are described by the same CDEs enabling the potential for data aggregation across independent groups.

USAGE:
These statistics from August 2010 represent usage of caDSR content in the C3D application used by CCR to support these cancer centers.

Sites: 200
Users: 1,124
Studies: 326
Patients: 13,514
CRFs: 910,071
CDEs used in C3D studies: 3,491
Individual data points (Responses): 150,341,554
Lab CRFs from caDSR download CDEs: 604,128
Data points from batch loaded CRFs: 124,480,823

These figures include the NCCCP RQRS and Breast Cancer Datamart (BCDM). For comparison, for October, 2010 the figures show 935,573 CRFs, an increase of 25,502.

NCI Cancer Therapy Evaluation Program (CTEP)

Project Sponsor: Dr. Jeff Abrams
Project Manager: —

The NCI Cancer Therapy Evaluation Program coordinates the largest, publicly funded oncology clinical trials organization in the world. With over 900 active trials enrolling annually 30,000 study participants, nearly 400 grants and cooperative agreements, and about 100 investigational new drugs (INDs), CTEP’s staff of physicians, scientists, pharmacists, nurses and regulatory specialists, assisted by government contractors, works diligently to assure the safe, efficient and ethical conduct of this complex research enterprise. CTEP-sponsored research spans phase 1-3 trials in all cancers and treatment modalities – chemotherapy, immunotherapy, radiation and surgery. All CTEP-reported trials are represented by data elements housed in the caDSR, with Case Report Forms to created to represent #NationalCancerInstituteCooperativeGroups trials and other trials.

USAGE: CTEP currently has 2,613 CRFs in caDSR with over 9,607 released and draft CDEs. See the #NCIFormBuilder description above for CTEP usage details.

NCI Division of Cancer Prevention (DCP)*

Project Sponsor: Dr. Leslie Ford
Project Manager: —

The Division of Cancer Prevention (DCP) is the primary unit of the National Cancer Institute devoted to cancer prevention research. DCP provides funding and administrative support to clinical and laboratory researchers, multidisciplinary teams, and collaborative, research-based networks. DCP also requires that all trials sponsored by the organization use variables that are included in the caDSR. Over the past decade nearly 1,300 data elements with 58 protocol forms and templates have been developed within the caDSR. DCP has also authored a significant number of the data standards vetted and approved by the caBIG® community over the past few years, including standards for the capture and reporting of:

- Address components
- Person name
- Person height
- Person weight

These data standards are widely reused across the NCI clinical research enterprise.

USAGE: DCP has 1,317 CDEs in caDSR. These are organized into categories such as Adverse Events Reporting, Protocol Deviation Notification, and types of conditions such as the 201 CDEs to collect for Barretts Esophagus, or the 152 CDEs for Bladder Cancer.

By using the caDSR and ISO/IEC 11179 to harmonize the descriptions of data to be collected in these studies, and designating the same CDEs to collect on different forms for the same data across these conditions, caDSR enables DCP to simplify data collection and training, as well as be assured of being able to aggregate data and perform comparative analysis across studies.

NCI Center for Cancer Research (CCR)
The Center for Cancer Research is the NCI's intramural research program. The Center (CCR) is home to more than 250 scientists and clinicians working in intramural research at NCI. CCR is organized into over 50 branches and laboratories, each one grouping scientists with complementary interests. CCR investigators are basic, clinical, and translational scientists who work together to advance our knowledge of cancer and AIDS and to develop new therapies against these diseases. CCR investigators collaborate with scientists at the more than 20 other Institutes and Centers of the National Institutes of Health (NIH), as well as with extramural scientists in academia and industry (as further discussed in the Appendix). CCR uses the NCI's C3D for study and trial data collection.

**USAGE:** As all C3D studies are described using CDEs from caDSR, so are all CCR studies. The CDEs are available electronically as an Excel download from a web browser-based UI, the CDE Browser, and then used to create forms in C3D. See Cancer Centralized Clinical Database (C3D) for details.

### Specialized Programs of Research Excellence (SPOREs)

**Project Sponsor:** —
**Project Manager:** —

Specialized Programs of Research Excellence (SPOREs) are funded through specialized center grants (P50s) that promote interdisciplinary research and move basic research findings from the laboratory to clinical settings, involving both cancer patients and populations at risk of cancer.

The outcome of interdisciplinary research is a bidirectional approach to translational research, moving laboratory discoveries to clinical settings or clinical observations to the laboratory environment. Laboratory and clinical scientists share the common goal of bringing novel ideas to clinical care settings that have the potential to reduce cancer incidence and mortality as well as improve survival and the quality of life. Since 2003 SPOREs studies predominantly for head and neck and lung cancer have been added to the caDSR. There are studies and CRFs for two Iloprost trials in the registry, as well as studies for the University of Colorado (GO studies for advanced lung cancer; Principal Investigator is Dr. Paul Bunn).

**USAGE:** As of October 2010, the SPORE program has registered 29 CRFs in caDSR from 704 CDEs. Of special note is the fact that the University of Colorado now has refined its protocol templates to the point that builds of new studies require essentially no creation of new elements; existing content can be rapidly used to build forms, increasing the interoperability and ability to aggregate collected data.

The Lung SPORE at Emory University Winship Cancer Institute has designated 91 CDEs to collect for Lung Cancer Patients.

### The National Marrow Donor Program® (NMDP) and our Be The Match FoundationSM Project

**Sponsor:** Dr. Roy Jones, MD Anderson; Dr. Doug Rizzo, Medical College of Wisconsin
**Project Manager:** Robinette Alley

The National Marrow Donor Program is a member of the Center for International Blood and Marrow Transplant Research. NMDP has located in Minneapolis and Milwaukee, with teams of data element curators who have been trained by preceptor NCI CBIIT staff. NMDP curators are using the caDSR Form Builder and Microsoft Excel add-in along with manual curation tools to register all the data collection forms and make them available via the caDSR web-based tools and programming interfaces. This organization began with a requirement to add nearly 100 CRFs to the registry; at the present time nearly all of the high volume forms have been completed, with CRFs designed for much more rarely seen malignancies slated for future curation.

Currently this group is evaluating the applicability of caAERS, an application designed by NCI CBIIT to document and report adverse events, for transplant research. They are also actively exploring the harmonization of their variables with those identified in the NCI Case Report Form Harmonization and Standardization initiative. The resulting set of variables for use in Transplant reporting will be vetted and evaluated as standards for the larger Transplant community. These activities have the strong support of Dr. Douglas Rizzo in Milwaukee, and Dr. Roy Jones at the MD Anderson Cancer Center. Dr. Jones is also a member of the caBIG® CTMS Steering Committee, and remains a strong proponent of the development of core data standards for clinical trial reporting.

**USAGE:** This group has identified for reuse or curated 1,535 of a planned 14,000 CDEs selected to use in their hematopoietic stem cell transplant (HSCT) program. NMDP utilizes AGNIS, a network based electronic forms system based on FormsNet to collect donor information. To harmonize the data across these 98 forms and render the CDEs programmatically accessible via caDSR APIs, NMDP has chosen caDSR to record their reusable data elements in ISO/IEC 11179 format, allowing them to develop sharable forms in caDSR Form Builder to build the collections of CDEs to match their existing varied and un harmonized FormsNet forms. The use of caDSR and careful planning by the NMDP curation team is allowing NMDP to harmonize fields across all donor program data collection forms and efforts (caBIG® 2010 Annual Meeting Poster 82).

### calintegrator

**Program Manager:** Juli Klemm

calintegrator is a web-based software package that allows researchers to set up custom, caBIG®-compatible web portals to conduct integrative research, without requiring programming experience. These portals bring together heterogeneous clinical, microarray and medical imaging data to enrich multidisciplinary research.

Using calintegrator, researchers can execute, save and share queries to identify and collect many types of data, combining clinical information with genetic and genomic data to enable multidimensional analysis. calintegrator uses caGrid analytical services such as GenePattern and BioConductor to perform analysis on the integrated study data, including clinical survival data.

calintegrator leverages the Cancer Data Standards Registry and Repository (caDSR) to map experimental data to well-defined datatypes and utilizes caGrid and Java client APIs to access data from caBIG® applications such as caArray and the National Biomedical Imaging Archive (NBJA). calintegrator is also integrated with caBio to perform queries on genes and pathways. NCI hosts an online version of the calintegrator application.
Usage: The caIntegrator has 362 CDEs described in caDSR, enabling users of the application to better understand the kind of data that is available in caIntegrator without having to look at the caIntegrator code. For example, the UML Model Browser can be used to see that there are 70 classes in the database, and that the field for “Class Comparison Analysis Adjustment Type MultipleComparisonAdjustmentType” with public ID and version 2529267v1. 0, with only 3 possible values in this field: “fdr”, “fwer” and “none”. The caDSR metadata provides the meaning of this data field, the datatype, representation and the meaning of each of the three values.

This information is then available to be accessed programmatically when designing the application, or at run time for the end user. An example of the details that are available for every CDE can be seen in these search results.

caB2B

Project Manager: Rakesh, Washington State University

cancer Bench-to-Beside (caB2B) is an open-source query tool that permits translational research scientists to search and combine data from virtually any caGrid data service.

The caB2B suite is composed of three core components: the Web Application, the Client Application and the Administrative Module. The caB2B Web Application provides query templates that leverage semantic metadata from caDSR to allow easy search and retrieval of microarray data (from caArray), imaging data (from the National Biomedical Imaging Archive (NBIA)), specimen data (from caTissue) and nanoparticle data (from caNanoLab) across the grid. Searches can be performed on selected locations using either form-based or keyword searches and data can be exported in the CSV format.

Using model metadata stored in caDSR to detect relationships between data classes, the caB2B Client Application is a thick Java application that enables advanced end users to create and execute queries across caGrid data services. The query component consists of a diagrammatic view that allows the user to create a directed acyclic graph of the query that is to be executed and also helps the user to connect two or more classes to be searched. Users can save the query and data returned may be saved in the form of a ‘virtual experiment.’ These data can be visualized using various graphical components.

The Administrative Module provides a graphical user interface for customizing a particular instance of caB2B. For example, an administrator can select models and service instances that may be queried, curate paths between classes in models, create user-friendly categories using two or more classes from a model, and create inter-model joins.

USAGE: caB2B utilizes caDSR CDEs and UML Model association information to support semantic queries and data aggregation across semantically and syntactically similar data on the grid.

The Cancer Genome Atlas (TCGA)

Letter: Dr. Kenna Shaw, Director, TCGA

TCGA teams work iteratively with NCI curator staff to identify disease-centric data element collections that use existing standards and extend standard content with TCGA domain expert definitions. Currently the TCGA project uses more than 1000 CDEs registered in the caDSR. This collaboration with the NCI caDSR team began in the spring of 2010, and continues to represent a major program commitment of caDSR resources to the TCGA program. The result has been a rich collection of elements that is heavily reused by other members of the caDSR extended community. Vocabulary needs are managed rapidly to allow the creation of new elements to proceed on an expedited timeline.

USAGE: The Cancer Genome Atlas has 636 CDEs registered in caDSR.

Imaging (In Vivo Imaging Workgroup, DICOM, ACRIN)

Project Sponsor: —
Project Manager: —

Imaging content is central to the oncology community and research. To this end, and based on early work with Cancer Imaging Program (CIP), the addition of data elements and models to the caDSR began in 2004. Early efforts to move legacy databases and applications quickly transitioned into the creation of new applications and intersections with the Imaging standards community. Content to reflect the Digital Imaging and Communications in Medicine or DICOM standard has been added to the caDSR using manual caCORE tools, and working collaboratively and iteratively with the imaging community.

A variety of applications as well as imaging variables used in clinical trials has raised the full set of imaging content in the registry to nearly 1400 variables. These variables are all annotated and referenced in a manner to maintain their links to the underlying DICOM standard. This task is facilitated by members of the American College of Radiology Imaging Network or ACRIN, who work within caBIG® and with curators to reinforce the need for registration of imaging standards.

C3D is one area in which the benefits of working with Imaging colleagues is most apparent. To the clinical research content required by a clinical trial we have added imaging content in caDSR, required by the NCI sponsors, resulting in a set of content that reuses clinical trials elements while extending out the set with most commonly ACRIN standards. This approach allows the rapid build of imaging clinical trials reporting their data through NCI mechanisms.

OpenMDR – Ohio State University

Poster: Rakesh Dhaval, Claiixo Melean, David Ervin, Philip Payne, PhD

USAGE: OpenMDR adopted ISO/IEC 11179 as the basis for describing data elements in caGrid services. openMDR allows model and application owners to search for reusable content from caDSR using Enterprise Architect and the MDR generator to create caGrid services with semantics based on CDEs (caBIG® 2010 Annual Meeting Poster 112).

DECAMAP - Data Element Curation & Mapping Platform

Project Sponsor: Dr. Chris Chute, Dr. Robert Freimuth, Dr. Jyotishman Pathak
"DECAMAP is a tool that allows PGRN researchers to harmonize their local data dictionaries to existing metadata and terminology standards such as the caDSR (Cancer Data Standards Registry and Repository), NCIT (NCI Thesaurus) and SNOMED-CT (Systematized Nomenclature of Medicine-Clinical Terms). This tool can be used to search/browse metadata related to different studies, create new study dictionaries and its related metadata, export metadata in excel format and assist investigators with dbGAP submission. DECAMAP is being created by the PHONT group in the PGRN network with the first version to be released July 2011. The overarching aim of our PGRN Ontology Network Resource (PHONT) is to enable codification of standardized phenotype definitions and relationships, in coordination with other established government-funded efforts. Because the standardized representation of phenotypes depends on standards, our first two aims focus on making vocabularies and ontologies (computationally formal vocabularies) about clinical and physiologic concepts accessible and usable to the PGRN community."

The focus of this research is on matching to the ISO 11179 DEC and Value Domain Permissible Values in caDSR to compare/string match values across different resources. The UI allows item by item searching. The user creates their own data dictionary and then uses that dictionary content to search in caDSR, NCIt, SNOMED and SDTM. An expert curator reviews the submitted data dictionary entries to find the best match. Based on sample studies, 69% matched to caDSR, 49% mapped to NCIt, lesser to SNOMED and SDTM.

http://informatics.mayo.edu/phont/index.php/Main_Page

**eMerge - Data Element Curation & Mapping Platform**


Dr. Dan Masys, Wang J, Kashyap S, Basford M, Li R, Dr. Jyotishman Pathak

Abstract: Systematic study of clinical phenotypes is important for a better understanding of the genetic basis of human diseases and more effective gene-based disease management. A key aspect in facilitating such studies requires standardized representation of the phenotype data using common data elements (CDEs) and controlled biomedical vocabularies. In this study, the authors analyzed how a limited subset of phenotypic data is amenable to common definition and standardized collection, as well as how their adoption in large-scale epidemiological and genome-wide studies can significantly facilitate cross-study analysis.

Conclusions: This study emphasizes the requirement for standardized representation of clinical research data using existing metadata and terminology resources and provides simple techniques and software for data element mapping using experiences from the eMERGE Network.


**caDSR Usage in NIH**

**National Heart Lung and Blood Institute (NHLBI)**

Project Sponsor: Dr. Jenny Larson, Dr. Douglas Rizzo

Project Manager: —

The National Heart, Lung and Blood Institute (NHLBI) began partnering with the NCI to add data elements to the caDSR in 2005 to reflect the variables used for a Family Blood Pressure Program (FBPP) study.

Working with NCI CBIIT curators, NHLBI achieved reuse of some existing elements, with registration of a significant amount of cardiology-specific content over the ensuing year. This is also the context into which the NMDP elements are being added.

**USAGE:** There are currently more than 4000 data elements registered in caDSR; a number of these are being used by oncology groups, demonstrating the ability to reuse variables across the research community regardless of the domain of expertise.

**National Institute of Dental and Craniofacial Research (NIDCR)**

Project Manager: Alice Birnbaum, Director of Biostatistics, Axio Research, LLC

The National Institute of Dental and Craniofacial Research (NIDCR) was one of the first non-cancer adopters of the caDSR as an attractive place for the registration of data elements. Since early 2005, a number of curation teams across the US have been trained by NCI CBIIT staff in the best practices for data element creation and maintenance.

**USAGE:** To date, more than 2300 data elements have been registered in the caDSR, and 81 Case REport forms, notably with the addition of valuable and unique dental-specific content. This organization is one of the groups driving the development of batch curation and registration functionality for the caDSR; the curators created a set of content that has been loaded using the caDSR Bulk Loader. We continue to capture the need for enhancements and emerging business requirements.

**National Institute of Child Health and Development (NICHD)**

Project Sponsor: Dr. Steven Hirschfeld

Project Manager: —

The National Institute of Child Health and Development is one of the newest non-oncology users of the caDSR. Beginning in November, 2009, a model to represent a Newborn Assessment was annotated and loaded into the caDSR. The resulting 107 data elements were reviewed and released for widespread community review.

**USAGE:** Currently a small team is being trained by NCI CBIIT staff to act as curators for additional content that is being planned for addition to the caDSR, as of 2011 114 CDEs have been created for use in this research domain, 7 have been reused from other Contexts.

**National Institute of Neurological Diseases and Stroke (NINDS)**
The National Institute of Neurological Diseases and Stroke (NINDS) has developed a set of data elements for use in clinical trials over the past several years. In the past year members of an NINDS CRO, KAI Research, recommended to NINDS that their core data elements may be candidates for addition to the caDSR. Based on their evaluation of the caDSR, a comparison analysis between NINDS content and NCI content was conducted. The result was in excess of 80% of core research variables across the two groups, again demonstrating the potential reuse of content regardless of the area of expertise. A small team of curators has been trained by NCI CBIIT, and curation of an identified core set of NINDS variables was slated to begin in August, 2010. After the core set is loaded, next steps will be determined.

**USAGE:** NINDS has reused and created 310+ CDEs found on 21 NINDS CRFs (123 reused CDEs) used to collect for neurological clinical research data.

**caDSR External Adoption**

NCI caDSR provides open access to all its content via publicly accessible browsers and representatives have spoken at a number of external conferences on the caDSR approach to data management and services as a means for sharing data. This has drawn the attention of a number of external groups and international parties looking to find data descriptions for clinical trial and research datasets as highlighted below.

**Biopathology Center, National’s Childrens Hospital**

Project Sponsor: Dave Billiter

Cooperative Group Specimen Banks are using caDSR content via caDSR API in their Group Banking Committee tooling to inspect and harmonize across different specimen banks. They use the caDSR API and the UML Model Browser and CDE Browser to identify CDEs that need to be harmonized.

**PhenX Toolkit**

The Toolkit provides standard measures related to complex diseases, phenotypic traits and environmental exposures. Use of PhenX measures facilitates combining data from a variety of studies, and makes it easy for investigators to expand a study design beyond the primary research focus. The data elements use to create these standard measures are based on 660 CDEs from caDSR.

Toolkit Users can:
- Search or Browse the Toolkit to review and select PhenX measures
- Request a Data Dictionary and Data Collection Worksheet for selected PhenX measures

https://www.phenxtoolkit.org/index.php

**GIATE**

Project Sponsor: The Antibody Society

Guidelines and Information About Therapy Experiments (GIATE) is a data standard for recording data about therapy experiments. It is being promoted by the Antibody Society, as a data standard for antibody therapy experiments but it has become apparent that the scope of materials used in antibodies as therapeutic agents, means that the data standard can equally be used to describe a wide range of therapeutics.

GIATE categorizes the information about therapy experiments into three classes, that is, information about the target, therapy agent and models used to test the therapy. Nested within these classes are additional classes describing different aspects and properties of the parent class. Therefore, a logical representation of GIATE would be a tree, with three main branches, which stems into smaller branches and leaves. The contents of the GIATE tree are downloaded from the caDSR, each concept contains a reference to the caDSR data element. http://www.genscript.com/giate-viewer/

**American Heart Association (AHA) and American College of Cardiology Foundation (ACCF)**

Project Sponsor: AHA, Duke

"The American College of Cardiology Foundation (ACCF) and the American Heart Association (AHA) support their members’ goal-to improve the prevention and care of cardiovascular diseases through professional education, research, and development of guidelines and standards and by fostering policy that supports optimal patient outcomes. The ACCF and AHA recognize the importance of the use of clinical data standards for patient management, assessment of outcomes, and conduct of research, and the importance of defining the processes and outcomes of clinical care, whether in randomized trials, observational studies, registries, or quality-improvement initiatives.” The AHA and ACCF have chosen the caDSR define and disseminate clinical data standards - sets of standardized data elements and corresponding definitions - to collect data relevant to cardiovascular conditions. http://circ.ahajournals.org/content/124/1/103.extract

**Winthrop P. Rockefeller Cancer Institute at the University of Arkansas for Medical Sciences**

Project Sponsor: Umit Topaloglu, PhD

The institute has decided to implement standard operating procedures (SOPs) requiring that all investigator led trials use CDEs from caDSR to conduct their research. The first project created 12 CRFs in the program's Clinical Data Management System (CDMS), using 72 CDEs, 38 of which were already in caDSR (50% reused). They created the remaining 33 CDEs using the Curation Tool. They have installed openMDR (OSU caDSR ISO/IEC 11179 compatible registry that can pull CDEs from caDSR for a local repository). The program plans to use this for ongoing collaborative research studies (caBIG® 2010 annual meeting Poster 80).
Union of Light Ion Centers in Europe (ULICE)

Project Sponsor: EU, WP Lead Dr. Bleddyn Jones, Manchester University, Oxford Computing Laboratory, Oxford UK

Informatics PIs: Professor Jim Davies, DPhil. Steve Harris, PhD

Project: WP7 Common database and grid infrastructures for improving and catalysing access to RI for the broad European community

ULICE is a 4-year project set up by 20 leading European research organizations, including 2 leading European industrial partners (Siemens and IBA), to respond to the need for greater access to hadron-therapy facilities for particle therapy research. Project coordinator is the Italian Research Infrastructure Facility CNAO (Milan).

WP7 is concerned in part with the compilation of a sufficient body of evidence to support future evidence-based evaluations of CPT – identifying new, preferred treatment protocols for specific cancers and ensuring that patients receive the best treatment.

By establishing a registry using key shared terminology and caDSR-type data elements to collect data, WP7 will collect information on every patient treated. However, such a registry will deliver the evidence only if the data gathered is comparable or interoperable. It is necessary to be able to reliably combine data recorded in different contexts, in different clinics, and in different health care systems.

Given the variety of (changing) practices and situations, it is unrealistic to expect that a single quality standard might be imposed for data collection (Bray and Parkin 2009) (Parkin and Bray 2009), so the registry has been set up to allow all data collected in the registry to be properly described, with detailed explanations of protocol for each clinical observation, and detailed definitions for each of the possible values recorded. These detailed explanations of the data will also be collected using caDSR-style common data elements. The ULICE project is in Year 1 of a 48 month project; it is aligning data elements, wherever possible, with caDSR. Currently 50% of the data elements to describe ULICE studies are caDSR based CDEs, as are 25% of the Baseline CDEs and 70% of the Followup CDEs.


Winthrop P. Rockefeller Cancer Institute, Cancer Control Core at the University of Arkansas for Medical Sciences; Breast Mammography

Project Sponsor: Umit Topaloglu, PhD

USAGE: The cancer institute is using caDSR CDEs and LimeSurvey, an open source tool, to recreate questionnaires for breast cancer mammography. The center recreates or reuses the survey CDEs in caDSR, then imports the CDEs into LimeSurvey. This ensures that the data are collected with high quality and common, standard data descriptions across studies. As a result of doing this work, the existing survey data was analyzed, corrected and standardized and can be reused for other studies (caBIG® 2010 Annual Meeting poster 81).

Oncology Patient Enrolment Network (OPEN) - Sponsor-CTEP/NCI

Project Manager: Mike Montello (Project officer)
Bioinformatics: Ravi Rajaram

The OPEN System has adopted the caCORE Application Programming Interface (API) and the Form Builder application, in addition to integrating and adapting the Common Security Model (CSM) components, as part of its architecture. OPEN uses the caDSR API to download the case report form (CRF) metadata from the caDSR. In addition, OPEN has adapted the CSM in such a way that the authorization feature of the CSM is used while authenticating the user against the National Cancer Institute – Cancer Therapy Evaluation Program (NCI CTEP) Identity and Access Management (IAM) system. The CSM user provisioning tool (UPT) is being used to provide instance and attribute level security.

Duke Translational Medicine Institute

Bioinformatics Sponsors: Dr. Jeffrey Ferranti, Patricia Gunter

M. Nahm, PhD, Associate Director, DTMI Biomedical Informatics
H. Shang, Director of Business Information Services, Duke Health Technology Solutions
Rob Califf, M.D., Vice Chancellor of Clinical Research, Duke University Health Systems
Asif Ahmad, CIO, Duke Health Technology Solutions
Dwight Smith, Director, Information Technology Application Development, Duke Health Technology Solutions

Dr. Ferranti at DTMI recognized a business problem in the lack of standardized clinical terminology across the enterprise, preventing meaningful sharing and reuse of the data – semantic interoperability.

Following NCI's lead for using controlled terminology as the basis for metadata descriptions (caDSR), the DIMI solution is to build a metadata registry that would become a foundation for future data management and data quality efforts. Currently Duke is considering caDSR open source as a possible solution versus building anew.

Silos of operational databases and data entry applications exist in Patient registration, Patient billing, Clinical systems – surgical, emergency, labs, and Disease registries in over 150 clinics with over 3 million HL7 transactions per day. To promote consistent semantics, DTMI would leverage controlled terminologies – LOINC, SNOMED-CT, ICD-9, CPT, and others, and the ISO/IEC 11179 standard for metadata registries and HL7 messaging standards. Like many institutions, DTMI data was collected on forms. The current repository collected these forms as a blob of data. DTMI used its 11179 registry to create structured descriptions of the data on these forms and used the metadata descriptions as a means by which to extract and transform the data into a data warehouse. The major difference between the NCI's caDSR and Duke's metadata registry is the terminology basis. NCI has used the NCI Thesaurus as the basis for describing data; DTMI is using concepts from the HL7 RIM as the basis for the metadata descriptions.

Taiwan Cancer Registry
ISO/IEC 11179 has been used to describe the cancer data set for the Taiwanese cancer registry. A paper entitled “Annotating Taiwan Cancer Registry to caDSR for International Interoperability” in “Future Wireless Networks and Information Systems”, 2012 explains that “It is very difficult to exchange and integrate the data among different cancer institutions, in order to discover useful biomedical information for the cancer community. We developed a cancer Biomedical Informatics Grid® (caBIG®) Silver level compliant cancer registry database system, called the gridTCR (grid of the Taiwan Cancer Registry), which integrates the demographic data, clinical history, pathology data, and clinical outcome data including treatment, recurrence and vital status from cancer registry databases in Taiwan. In this manuscript, we will be developing the common data elements using vocabulary standards, ontology and semantic modeling methodology. The Taiwan Cancer Common Data Element Project (TCCDEP) developed 40 data elements to annotate the cancer registry data collected. The aim of this project is to create a core set of data elements for annotating the cancer registry data and achieve the interoperability over the caBIG community. We describe the process required to develop the model, the caDSR CDEs, and the results of the modeling effort. We address difficulties we encountered and modifications for solution. Currently, the Taiwan cancer registry CDEs are released and available in CDE browser for reusing. Furthermore, we will extend our CDEs to daily clinical practice and trials, along with how the methods were used to fully implemented in hospitals and cancer research centers in Taiwan.”

**Children’s Oncology Group (COG) and National Childhood Cancers Foundation**

John Deardurff, Nationwide Childrens Hospital

The COG and National Childhood Cancers Foundation are using caDSR APIs to retrieve CDEs for development of studies using common data descriptors.

**University of Miami Ontology Modeling**

Yevs R. Jean-Mary

Using UML models registered in caDSR and related NCIt concepts, the University of Miami uses semantic web technologies including OWL to develop executable queries against NCIt concepts that can return instances of caGrid Data, (caBIG® 2010 annual meeting poster 87).

**Center for International Blood and Marrow Transplant Research (CIBMTR)**

Kirt Schapner, Robinette Aley, Barb Kramer, Dr. Douglass Rizzo

This project groups similar data elements using ISO/IEC 11179 structures.

The Center for International Blood and Marrow Transplant Research (CIBMTR) is collecting data on 45 forms over 5,000 data points. The program uses clustering techniques to analyze the parts of CDEs (Data Element Concept, Object Class and Property and Value Domain concept associations) to discover similar data points. These will be used to help create an HSCT-specific database model in the future (caBIG® 2010 annual meeting poster 95).

**The Human Studies Database (HSDB)**

UCSF (Leader): Ida Sim (Project Leader), Simona Carini, Rob Wynden

The project goals are to:

- Enhance national clinical and translational research capability: As a computable database with detailed scientific information about past and ongoing human studies, HSDB will be a critical infrastructure for large-scale analysis and reuse of human studies data.
- Ensure standardized computable descriptions of human studies through use of ontologies and controlled clinical vocabularies. The semantic foundation for HSDB is the Ontology of Clinical Research, with mappings to and integration with other relevant models and vocabularies (for example, BRIDG, SNOMED). By adopting ISO/IEC 11179 as the framework for representing the meaning of data, this semantic foundation is independent of the particular technology or architecture used for implementing data sharing.
- Provide a set of tools for institutions to contribute their human studies data to HSDB and to share this data over caGrid. We are focusing first on federating descriptions of study designs before federating study results. Definition and reuse of openMDR and caDSR CDEs will help to ensure consistency in the data collected across these study designs.

Current project members are primarily from Clinical and Translational Science Award (CTSA) institutions, the HSDB project is open and other participating institutions are:

- Duke University: Meredith Nahm, Swati Chakraborty
- Columbia University: Suzanne Bakken
- Johns Hopkins: Harold P. Lehmann
- Mayo Clinic: Chris Chute
- The Rockefeller University: Ed Barbour, Shamim A. Mollah, Knut M. Wittkowski
- Stanford: Samson Tu
- UCSF Davis: Davera Gabriel, Hien Nguyen
- University of Colorado: Jessica Bondy
- University of Manchester, UK: Alan Rector
- UT Health Science Center San Antonio: Brad H. Pollock
- UT Southwestern: Herbert K. Hagler, Richard H. Scheuermann
- University of Washington: Jim Brinkley, Todd Detweiller
- Washington University St. Louis: Rakesh Nagarajan, Jahangheer Shaik

This consortium has adopted the caDSR-compatible openMDR as the basis for registering semantic descriptions of data in the HSDB database.
Towards Unambiguous formal descriptions of cancer therapy experiments

Poster (and a paper reference TBD), authors: Alejandra González-Beltrán, PhD et al, Wolfson Institute of Biomedical Research and Department of Computer Science and Cancer Institute, University College London

caDSR CDEs and new CDEs based on ISO/IEC 11179 structure were developed for the Guidelines for Information about Therapy Experiments (GIATE). The CDEs are annotated with concepts from the NCI Thesaurus. Though GAITE was developed independently of NCI Thesaurus, by creating the structured CDEs, using OWL and following a semantic web and linked data approach, an unambiguous and formal description of GIATE was developed from the CDE structures and used to compare GAITE to the ontologies in NCI Thesaurus. “This matching will facilitate interoperability between GIATE compliant knowledge bases and caBIG® services.” (caBIG® 2010 annual meeting poster 78)

Ontology based queries for caGrid Infrastructure

Poster, authors: Alejandra González-Beltrán, PhD, Ben Tagger, Eng.D., Anthony Finkelstein, B.Eng, M.Sc., Ph.D., C.Eng, FIET, CTP, FBCS et al, Wolfson Institute of Biomedical Research and Department of Computer Science and Cancer Institute, University College London

Using OWL as a formal language for representing knowledge, this team was able to leverage caDSR mappings between UML Models and NCI Thesaurus concepts and ontology which serves as a conceptual unified view of the data services. This project developed a service for caGrid that provides methods to take caDSR registered models and turn them into OWL. The project has developed a query process in light of OWL2 profiles (caBIG® Annual Meeting poster 79).

Information Management Services, Inc. (IMS) BIS, NAACCR and DCCPS/caSEER program

The program provides NIH, pharmaceutical, academic and other research organizations with biomedical computing support including:

- Web and Application Development
- caBIG® tool adoption and enhancement support
- Repository Management Systems
- RISMA Compliant Computer Center Hosting
- Biomedical Research Data Management and Study Research Support
- Statistical Analysis and Reporting
IMS supports the development of CDEs for caDSR for the Person Age standard and the NAACCR CDEs. The program developed data models for seven data services for DCCPS/caSEER Program for sharing through the grid and has registered the CDEs in the caDSR. The program’s databases contain over 6 million records that can be searched and summarized to produce tables, graphs, and geographic maps for caSEER. Applications developed by IM are designed to extract and publish specimen data using caBIG® CDEs and the caBIG® Common Business Model.

The CTSA Health Ontology Mapper

This application allows users to take local data and map it to ontology codes using caDSR and LexEVS. The application queries against the NCI LexEVS api server, as well as caDSR and caGRID. Further information can be found on the Health Ontology Mapper website.

A Growable Network Information System (AGNIS)

MDACC: Dr. Roy Jones, Charles Martinez

AGNIS, or A Growable Network Information System, began as an idea to transmit hematopoietic stem cell transplant data between organizations but developed into a standards-based communications model. AGNIS was originally intended to be built on an existing messaging system. However, analysis of the National Institutes of Health (NIH) caBIG® project produced a strong case for using caBIG® tools, as caBIG® is becoming a recognized leader in creating standards for grid computing and data definition. The mission to connect the cancer community to accelerate research discoveries and improve patient outcomes fits the purpose of AGNIS, which is the implementation of clinical data exchange across the HSCT community to decrease the time it takes for patient follow-up data to be available for research.

AGNIS works by acting as a translator between two centers. When organizations attempt to communicate their data to one another, they must first massage the data into a format that the other organization can recognize. The translation mechanism is actually standardized common data elements (CDEs) curated and stored on the NIH's publicly accessible caDSR. Due to the reusable nature of elements in caDSR, everyone benefits as more terms are created.

USAGE: The AGNIS network uses the CDE Browser, Curation Tool and caDSR APIs to create and access standard CDEs.

cancerGrid, cgMDR UK

Project Sponsors: Peter McCallum, Jim Davies, Steve Harris

cancergrid is an initiative involving scientists at the Universities of Oxford and Cambridge, working together to reduce the cost of clinical research, and to increase its value through effective data sharing. It is building upon the success of a four-year project funded by the UK Medical Research Council, to address a wider range of scientific goals, with support from Microsoft Research.

The cancergrid team have developed standards and tools for the automatic production of the systems needed to support clinical studies and translational research. The vision:

- the researcher creates a model of their study or dataset, based upon standard templates, using a simple study designer tool
- the software artifacts - forms and services - needed to run the study, or interact with the dataset, are then produced automatically from the model

Along with a meta-model (or model template) for clinical studies, based upon the CONSORT statement, and work on the classification and registration of clinical trials, the project has produced an ISO11179-compliant metadata registry, a semantically-aware trials design/ management system, and a toolkit for clinical data transformation and integration.

These models and software applications have been tested through initial deployment on a small number of clinical studies in Oxford and Cambridge. The metadata registry software has been adopted for use within the US caBIG initiative, and is being evaluated by a number of organisations within the UK. Development continues at Oxford, with clinical collaboration in Cambridge and London, and technical collaboration with the caBIG team in the US.

http://www.cancergrid.org/

USAGE: cgMDR was built to be compatible with caDSR, including extensions to ISO 11179 to reference concepts as the grounding for semantics. They have developed an import that can take caDSR CDEs downloaded into an XML file and import then into local cgMDR installations. They have included the "Standard" CDEs in their repository as well as including BRIDG CDEs in their registry download so that local installations may more easily reuse /harmonize with NCI standards as much as possible.

CDC, PHIN Reportable Conditions and Common Core Data Elements

Project Sponsors: Michael Pray, Sundak Gunesan, Catherine Staes

Public health data elements and value sets described in over 90 CSTE position statements which are the foundation building blocks of Public Health Case Reporting (PHCR), ELR and Case Notification. A project is underway collaboratively distribute these data elements and value sets using the existing tools and applications at the National Cancer Institute through the caDSR CDE browser and CDC Vocabulary Server (VADS).

Related Software Engineering Research Highlights

"Quality evaluation of cancer study Common Data Elements using the UMLS Semantic Network"

Authors: Guoqian Jiang, Harold R. Solbrig, Christopher G. Chute

This paper discusses the use of UMLS SemNet to discover disjointness of semantic network types in caSDR Object Class (OC) and Property (Prop) as a QA mechanism for validating the OC and property in the Data Element Concept. This could be the basis for a new collaborative curation platform, where these kinds of rules could be incorporated as OWL reasoning for automated validation.

It also discusses using UMLS SemNet as an alternate classification scheme for browsing content in caDSR.

Appendix A - Statistics

Below are a few statistics on some key caDSR resources.

Data Element Descriptions

44,682 Community CDEs across all Contexts

- 24,749 with Workflow Status of "Released"
- 26,939 used in caBIG®
- 10,462 used by CTEP
- 2,272 in CCR implementations
- 3,491 used in C3D studies
- 628 by the Specialized Program of Research Excellence (SPOREs Program)
- 1,373 used by DCP
- 129 used by EDRN
- 288 used by Cancer Imaging Program (CIP)
- 4,007 used by National Heart, Lung and Blood Institute (NHLBI), including National Marrow Donor Program
- 2,339 used by National Institute of Dental and Craniofacial Research (NIDCR)
- 114 used by National Institute of Child Health and Development (NICHD)
- 310 used by National Institute of Neurological Disorders and Stroke (NINDS)
- 1,112 used by Population Sciences and Cancer Control (PS&CC)

As seen in the first chart below, the trend in total numbers of CDEs has been increasing since the project's inception in 2002, with 100 to thousands of draft new CDEs added with each new UML Model and with each new adopter's project (NCI, NIH, External).

While the total number of elements in the registry is increasing, the focus in 2010 has been on harmonization and releasing CDEs that are used, vetted, or both by the community and on identifying and retiring unused content as depicted in the second table below. Retiring unused content will make it easier for new users to find good reusable content.

Chart showing trends in total numbers of CDEs
Chart showing effect of harmonization and retirement of unused content

**Case Report Forms and Surveys**

3,328 Forms created by the Community

- 2,923 CRFs in CTEP used in 611 trials
- 152 in CCR used in 322 Active Trials
- 102 in NHLBI
- 99 in caBIG®
- 92 in NIDCR
- 40 in DCP
- 29 in SPOREs
- 910,071 CRFs created in C3D from caDSR CDEs

**UML Models**

144 Models registered by the Community

- 130 with a Released Status
- NICHD (1)
- Newborn Examinations
- NCI Population Sciences and Cancer Control (5)
- NHIS2-005, caSEER, HINTS2005, SNP500Cancer, GridEnabled Measures
caDSR Tools and Browser

The statistics below only cover the tools most commonly used by the community.

CDE Browser (for the month of June 2011)

544 unique visitors to the site

Average number of unique visitors per day:


Unique visitors past 3 months: Apr 550, May 554, June 544

The top 10 non-NIH domains accessing the CDE Browser were:


Curation Tool (as of Feb 2011)

129 unique users visited the site this month.

Curation Tool was linked to by 119 distinct web sites.

Unique visitors past 3 months: Nov 108, Dec 111, Jan 115
(Compared with 2010 Nov 539, Dec 430, Jan 490 --)

Trend Year/Year - 19% Decrease, Average of 117 unique visitors for past 3 months (144 same period in 2009-10).

The top non-NIH Domains using Curation Tool were:

- verizon.net, nmdp.org, saic.net, comcastbusiness.net, direcpc.net, scenpro.com

Form Builder (as of Feb 2011)

101 unique visitors to Form Builder this month.

Form Builder was linked to by 31 distinct web sites.

Unique Visitors the past 3 months: Nov 83, Dec 75, Jan 101

Trend Year/Year - 12% Decrease, Average of 86 unique visitors for past 3 months (97 same period in 2009-10).

For the month, the top 5 non-NIH domains using Form Builder were:

- nmdp.org, Comcast.net, saic.com, verizon.net, direcpc.com
UML Model Browser (as of Feb 2011)
104 unique users visited the site this month.
UML Model Browser was linked to by Not Available distinct web sites.
Unique visitors past 3 months: Nov 99, Dec 96, Jan 104
Trend Year/Years - 22% Decrease, Average 99 unique visitors for past 3 months (138 same period 2009-10)
For the month, the top 5 non-NIH domains using the site were:
• cox.net, duke.edu, nmdp.org, scenpro.com, uams.edu

caDSR API services (as of June 2011)
161 unique users visited the caDSR API site this month.
caDSR API was linked to by 17 distinct web sites.
The top NIH referring URL was the Freestyle API, see statistics below.
Unique visitors past 3 months: Nov 531, Dec 163, and Jan 63
The top 10 non-NIH domains accessing caDSR API were:

Freestyle services (as of Feb 2011)
52 unique users visited the site.
Freestyle was linked to by 12 distinct web sites.
Unique visitors past 3 months: Nov 33, Dec 33, Jan 43
Trend – Freestyle was incorporated in SIW for searching for existing caDSR content when annotating models. It is currently used by caIntegrator.
Average 43 unique visitors per month.
For the month, the top 5 non-NIH domains using the site were:
• UNKNOWN (4,849 accesses), saic.com, cox.net, gatech.edu, as12448.com

Appendix B - Support for Clinical Trials

<table>
<thead>
<tr>
<th>Cancer Centers Supported by caDSR CDE development for clinical trials</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abramson Cancer Center of the University of Pennsylvania</td>
</tr>
<tr>
<td>Albert Einstein Medical Center</td>
</tr>
<tr>
<td>ACRIN</td>
</tr>
<tr>
<td>Arizona Cancer Center</td>
</tr>
<tr>
<td>UCI / Chao Family Comprehensive Cancer Center</td>
</tr>
<tr>
<td>Duke University</td>
</tr>
<tr>
<td>Massachusetts General Hospital/Dana Farber</td>
</tr>
<tr>
<td>Georgetown / Lombardi Medical Center</td>
</tr>
<tr>
<td>Johns Hopkins Medical Institute</td>
</tr>
<tr>
<td>MD Anderson Cancer Network</td>
</tr>
<tr>
<td>National Cancer Institute Center for Cancer Research</td>
</tr>
<tr>
<td>National Cancer Institute Cancer Diagnosis Program</td>
</tr>
<tr>
<td>Northwestern University</td>
</tr>
<tr>
<td>Oregon Health &amp; Science University</td>
</tr>
<tr>
<td>University of Colorado Health Sciences Center</td>
</tr>
<tr>
<td>St. Joseph Hospital of Orange</td>
</tr>
<tr>
<td>St. Joseph Hospital / Candler</td>
</tr>
</tbody>
</table>
## Appendix C - Models

Models with CDEs registered in caDSR, listed alphabetically and by Version within NCI program/Context owner:

Some of these models are retired. Model details are accessible via the UML Model Browser.

### caBIG® (NCI cancer Biomedical Informatics Grid Data Services)

- AIM 1.0
- AIM 1.5
- AIM 2.0
- AIM 3.0
- Bioconductor 1.0
- BiospecimenCoreResource 1.0
- BRIDG 1.0
- BRIDG 2.1
- C3D Connector 1.5
- C3PR 1.0
- C3PR 2.0
- C3PR 2.8
- caAERS 1.0
- caAERS 2.0
- caArray 2.0
- caArray 2.1
- caArray Internal 2.4
- caArray_1.1
- caBIO 4.0
- caBIO 4.1
- caBIO 4.2
- caBIO 4.3
- caCORRECT 1.0
- CAD Markup 1.0
- CAD Order 1.0
- caDSR 4.0
- caElmir 1.0
- caElmir 2.0
- caFE Server 2.0
- caGrid_Metadata_Models 1.0
- caIntegrator 2.0
- caIntegrator 2.1
- Caisis 3.5
- caLIMS2 1.1
- caMOD 2.5
- caNano 1.0
- caNanoLab 1.0
- caNanoLab 1.4
- caNanoLab 1.5
- Cancer Molecular Pages 1.0
- CAP Cancer Checklists 1.0
- Cardiovascular Model 1.0
- caTIES 1.0
- caTIES 2.0
- caTISSUE CAE 1.2
- caTISSUE Core 1.0
- caTissue_Core 1.1
- caTissue_Core_1_2
- caTissue_Core_caArray 1.0
- caTissue_Suite 1.0
- caTissue_Suite1_1 1.1
- caTissue_Suite1_2 1.2
- caTRIP Annotation Engine 1.0
- caTRIP Tumor Registry 1.0
- caXchange 1.0
PS&CC (NCI Population Sciences & Cancer Control)

caSEER 1.0
CESD 1.0
HINTSSurvey 1.0
NHIS 2005 1.0
SNP500Cancer 1.0

caCORE (NCI Core Infrastructure)

caCORE 3.0
cacORE 3.1
cacORE 3.2
MDR Object Cart 1.0
MicroArray Gene Expression Object Model (Mage-OM) 1.0
Appendix D - Publications

Publications citing caDSR and CDEs


5. Alejandro González-Beltrán, Ben Tagger, Anthony Finchkeilstein, "*Federated Ontology-based Queries over Cancer Data*". Accepted for publication BMC Bioinformatics


18. Alejandro González Beltrán, Anthony Finchkeilstein, J Max Wilkinson, Jeff Kramer "Domain Concept-Based Queries for Cancer Research Data Sources " In the proceedings of the 22nd IEEE International Symposium on Computer-Based Medical Systems 2009 (CBMS 2009). Special track on HealthGrid Computing - Applications to Biomedical Research and Healthcare


21. Davies, J., Gibbons, J., Harris, S., Warzel D. "Evolving Health Informatics Semantic Frameworks and Metadata-Driven Architectures", Position paper in proceedings and session presented at Microsoft eScience Conference, (December 2008), Indiana, USA


**Abstracts, Posters and Talks**


24. Yu Rang Park, MS and Ju Han Kim, M.D., Ph.D."Metadata registry and management system based on ISO 11179 for cancer clinical trials information system", *AMIA Annual Symposium*, 2006


Help Downloading Files
For help accessing PDF, RTF, MS Word, Excel, PowerPoint, audio, or video files on this wiki, go to Help Downloading Files.