2 - caDSR Infrastructure and Toolset Usage

2 - caDSR Infrastructure and Toolset Usage

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. It is comprised of a central database and a set of web based tools and software interfaces (APIs).

This section includes the following topics.

- Overview
- NCI caDSR CDE Browser
  - CDE Browser Description and Usage
  - CDE Collections as Prepackaged 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

Overview

Until 2010, all data, (30K+ "Released" data elements, 44k+ total), and information models (198 including multiple versions of some models) for all systems and applications in caBIG® were required to have the meaning coded with controlled terminology, and recorded in the caDSR via UML Models. The high number of data elements in caDSR represent the many and varied ways in which similar data are collected and stored in databases throughout the community and provides a means by which data transformation programs can be written to enable data to be combined, or to design new applications that can reuse existing data.

By exposing these details in a common, structured way using ISO/IEC 11179, caDSR tools enable the construction of 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 details of the data elements used in applications, the registry also records the meaning in reusable semantic components that provide the basis for detecting conceptual equivalence across data elements and thus improves the potential for data aggregation.

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.

Adoption/adaption includes development of an open source XML-based version of caDSR by the University of Oxford Computing Lab, cgMDR, to support the cancerGrid's clinical trails and other clinical and population studies in the UK. caDSR CDEs were downloaded, decomposed and imported into this registry to form the basis of new content development, reusing many of the NCI's CDEs in UK trials. 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 CDEs during UML Model design, generating caGrid compatible services for deployment on the grid. OSU has also used this registry, named openMDR, to support its Clinical and Translational Science Award (CTSA) program and another consortium involved in collecting human studies, HSDB.

caDSR usage can be measured by the large number of communities and applications that have registered their data elements and models in caDSR. It enables sharing of these descriptions among data managers and software developers through CDE Browser downloads and the use of caDSR APIs, including an HTTP API that can be used to browse and download content in HTML or XML.

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

CDE Browser Description and Usage

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.

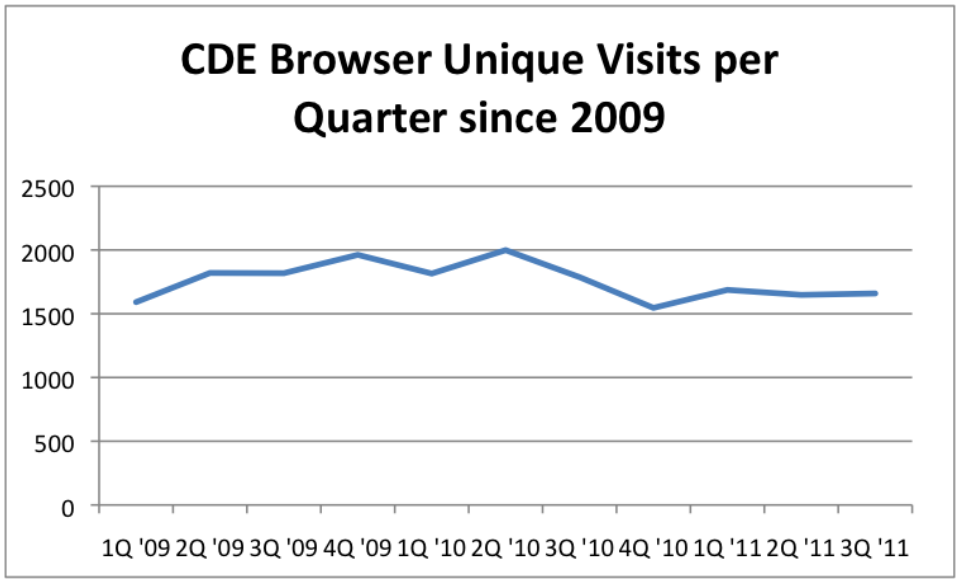
- Search for data elements by NCI Thesaurus Concept, Value Domain Permissible Value, Classifications or simple text searches in the names and definitions of the elements in caDSR.
- Search for classes of information in caDSR and find reusable CDEs. For example, advanced search panel can be used to search for the Property "Email Address" and retrieve the 43 that describe different types of Email addresses such as "Organizational Unit Email Address Text", "Person

Email Address Text", "Clinical Trial Participant Email Address Text" and "Investigator Email Address Text". The reuse of "Email Address" is recorded in the caDSR so that developers can detect the similarity between data elements and enable them to discover and potentially combine data from different sources.
- Using the Shopping Cart, create a customized set of CDEs for exporting in XML or Excel format or sharing with other applications such as Form Builder.
- The CDE Browser publishes its DTD to export/download data elements in XML.
- Access the Tool at https://cdebrowser.nci.nih.gov/CDEBrowser/

USAGE: An average of 47 unique users visit the CDE Browser each business 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.

The following chart shows the number of unique visitors, per quarter, to the CDE Browser from the 1st quarter of 2009 through the 3rd quarter of 2011.



The average number of visits began to rise in 2009, potentially due to the caBIG® program reaching maturity and that all data services had to have a UML model registered in caDSR, the CDE Browser being the primary means by which users could review and download the elements in their models after loading. The level of unique visits has stabilized so far in 2011 to a level equivalent to early 2009.

CDE Collections as Prepackaged File Downloads

Several collections of CDEs have been downloaded in Excel and XML 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 TCGA, NACCAR, SEER and BRIDG. Collections are posted to the caDSR Wiki on the "caDSR Hosted Data Standards, Downloads, and Transformation Utilities" page

NCI caDSR Form Builder

Form Builder allows users to share data element descriptions across multiple forms to ensure that data will be comparable. It helps users organize CDEs that replicate the content of Case Report Forms. These forms can include modules grouping CDEs together that can be copied from one form to another. Forms can include complex behavior such as skip patterns where the answer to a question determines the next question to be asked. Using the CDE Browser, 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.

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
- Classify theh CDEs on a 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
- Edit, Save a Forms shopping cart for storing collections of form to export to other systems through the Object Cart API

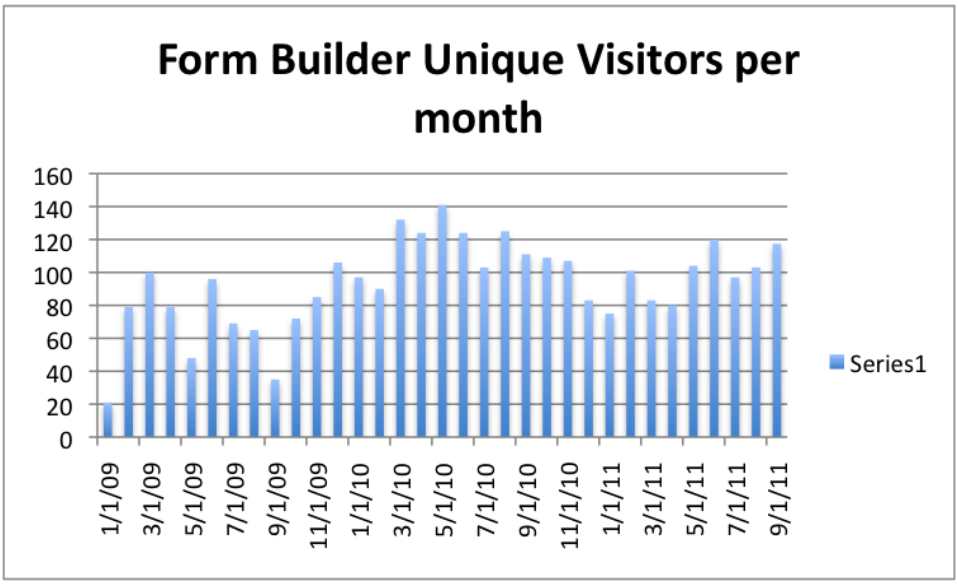
Access the Tool at https://formbuilder.nci.nih.gov

USAGE: As of October 2011, an average day there are 6 curators using Form Builder creating or modifying forms. There are 3,288 forms in caDSR, 466 Protocols (a 14% increase of 2010), and over 2,900 forms in the CTEP context, caDSR's largest single user (a 12% increase over 2010). CTEP Case Report Forms (CRFs) are grouped to 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.

The following chart shows the number of unique visitors, per month, to Form Builder from the 1st quarter of 2009 to the 3rd quarter of 2011.



NCI caDSR Curation Tool

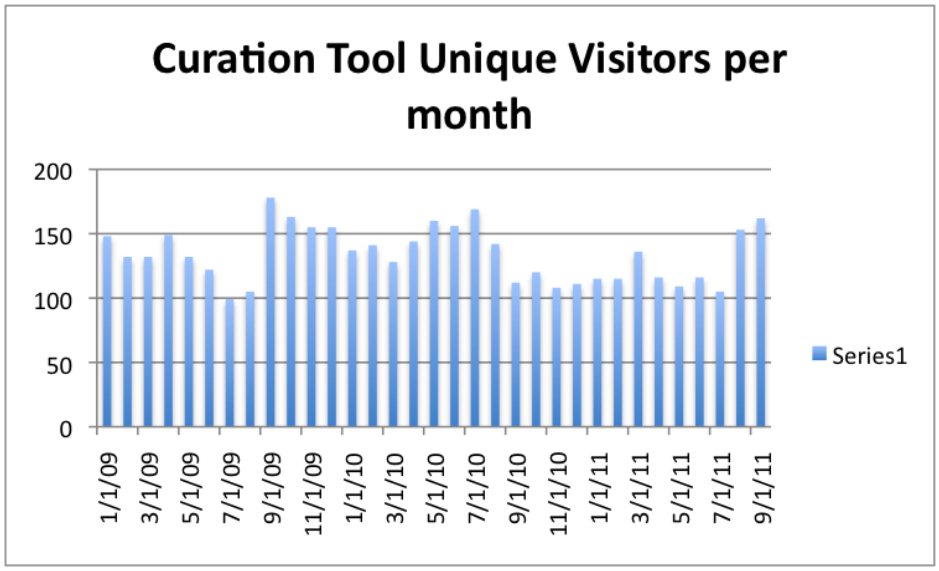
The CDE Curation Tool supports creation and editing of all the primary semantic descriptions for Data Elements used by the community in data repositories or application software. It is intended for use by Context Administrators. This tool's features facilitate the use of the CBIIT Enterprise Vocabulary 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.

Access the Tool at https://cdecurate.nci.nih.gov

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.

The following chart shows the number of unique visitors, per month. to the CDE Curation Tool from January 2009 to September of 2011.



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 is 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.

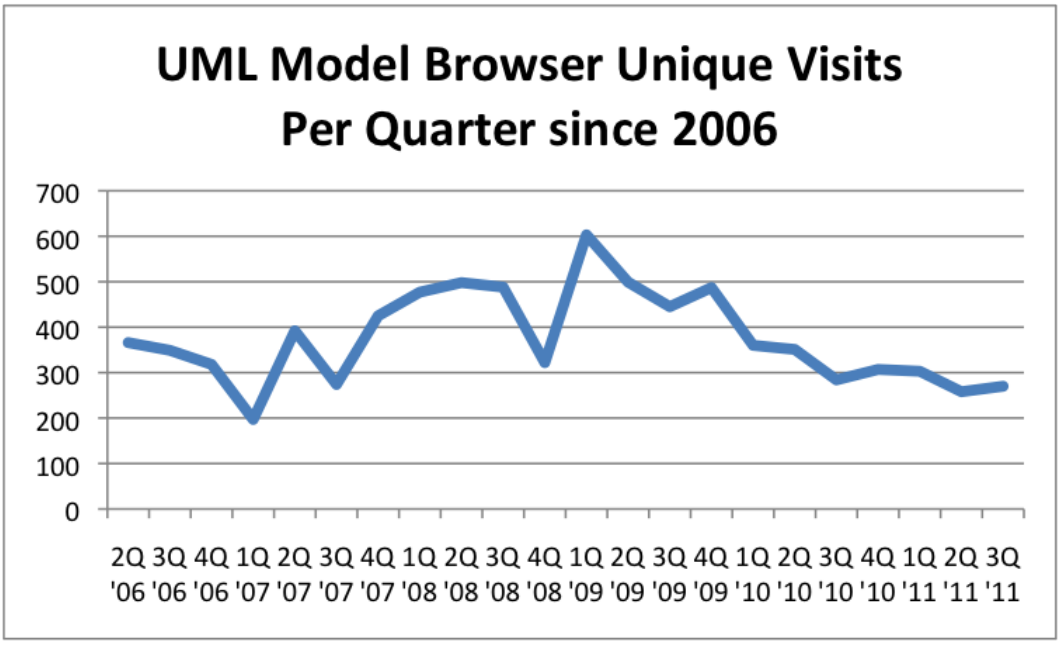
NCI caDSR UML Model Browser

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

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.

The chart shows the number of unique visitors, per quarter, to the UML Model Browser from the 2nd quarter of 2006 to September of 2011.



UML Model Browser usage peaked with loading of models by the caBIG® community during 2008 and 2009, and has leveled-off slightly lower than when the tool was first introduced.

The tool is useful for viewing the elements in a specific model, and we anticipate that with more emphasis on reusing content from existing models, the average number of visitors using the model browser will remain the same or increase slightly.

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

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.

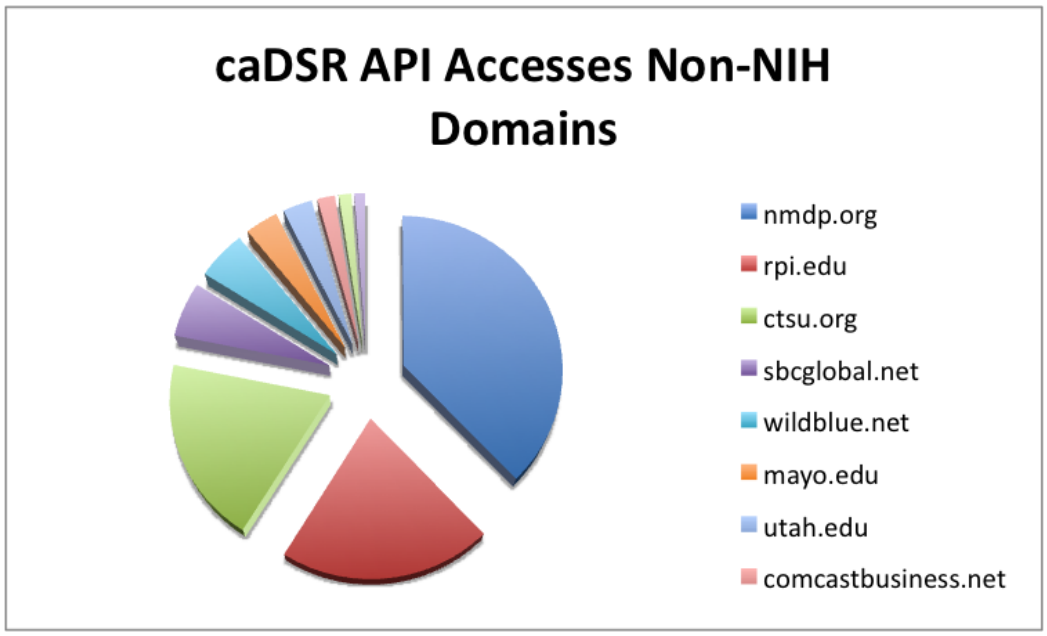
caDSR API allows users to access caDSR content by using a web browser to navigate the caDSR domain model and returns results in HTML or XML. A caDSR Java API provides a set of methods that can be used to retrieve content as XML documents. According to NCI statistics from Wusage, from October 2009 to October 2010, 628 unique visitors came to the site and accessed over 25M documents. The number of pages is slightly inflated sue 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.

Access the HTTP caDSR Domain Class Browser at http://cadsrapi.nci.nih.gov/cadsrapi40/

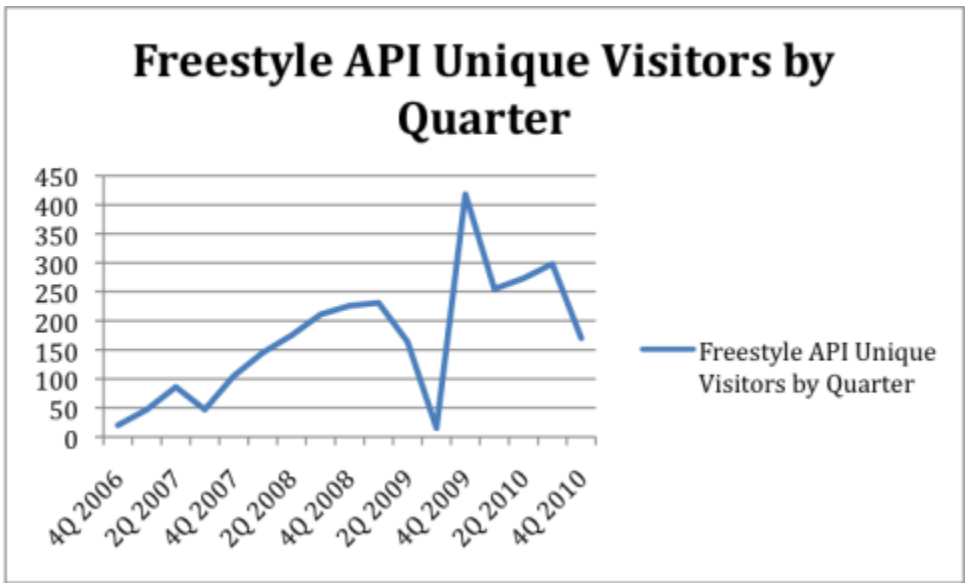
Freestyle API is another caDSR interface that provides access to content both via a web browser and application programming interface. From October 2009 to October 2010, 328 unique visitors came to the Freestyle API site and accessed over 9.6k documents. The Freestyle API uses the caDSR Domain API to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

Access the Tool at http://freestyle.nci.nih.gov

The chart that follows shows the comparative number of caDSR pages delivered through the caDSR API for various organizations during the period from 1st quarter 2009 to the 3rd quarter 2011.



The next chart shows the number of unique visitors to the Freestyle API during the period from 4th quarter 2006 through the 4th quarter 2010.



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.

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 XMI 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 (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, with many of the CDEs reused across models. For example, if the CDEs were evenly distributed across all models, each model would use only 133 CDEs. 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 Appendix C - Models Registered in caDSR for a complete list of models with CDEs in caDSR.

The chart that follows shows the comparative number of models processed through the SIW and loaded to caDSR via the UML Model Loader, by year, from 2005 through 2010.