CAMEO: A tool for secure aggregation of healthcare data

Research Area

This application addresses broad Challenge Area (10) Information Technology for Processing Health Care Data and specific Challenge Topic 10-RR-101: Information Technology Demonstration Projects Facilitating Secondary Use of Healthcare Data for Research.

Challenge and Potential Impact

As stated in the Challenge Topic, the analysis of aggregate, anonymous healthcare data, stored in electronic systems, has tremendous potential to positively impact healthcare delivery. However, significant technical challenges currently exist in acquiring and making primary healthcare data readily available for use in research, so that it is secure and meaningfully organized for searching, ease of access, and collaboration. In order to address this gap, various information technology projects have been initiated to facilitate secondary use of healthcare data in research. Recent projects include:

The University of Texas Health Science Center at Houston, is developing an integration platform for clinical research, built using semantic web technologies, that will integrate and classify biomedical data from disparate sources[i].

caTRIP[ii], the Cancer Translational Research Informatics Platform, has been developed at Duke University, using the caGrid infrastructure, as a platform for the aggregation of clinical and molecular data for use in research. caTRIP offers considerable functionality for conducting clinical research focused on cancer using sophisticated tools. The caTRIP project, in particular, highlights the unmet need for a system which is capable of aggregating clinical data for research "in a repository that is user-friendly, easily accessible, as well as compliant with regulatory requirements of privacy and security".

Both systems face the common challenge of acquiring data from point of care, and other source systems, in a structured and secure fashion. Both address this challenge by requiring the source systems to adhere to interoperability standards that are specific to the research platform (e.g., caGrid), and that may not already be present in the source systems. The UT project is in the relatively early stages of development, but recognizes the need for "services that enable provisioning and integration of datasets". They do not describe in detail how these services will be implemented, however it is fair to assume that individual services will be built for each source system to be integrated. Access to these services will be managed by an "authorization and control model" that spans all services and data handled by the system. Data is made available to caTRIP via the caGrid infrastructure. caGrid compatibility requires semantic interoperability[iii], which for a system that has not been designed for compatibility from the beginning, is a non-trivial process[iv].

HL7 standards are the most widely used interoperability specifications for healthcare data in the United States. More that 90% of all healthcare software vendors participate in the HL7 organization[v], and over 95% of all American hospitals and ambulatory facilities are using HL7 standards[vi]. However, despite the prevalence of systems capable of producing HL7 compliant data, no system exists to aggregate data in a comprehensive manner from these systems for clinical research. There is considerable value in providing a tool that implements a comprehensive mechanism for securing, indexing, and redistributing healthcare data represented as HL7 messages. Such a tool provides for relatively simple enablement of the acquisition of data from the vast majority of existing healthcare management systems by leveraging those systems existing capabilities.

To bridge this gap, we propose to build CAMEO: A scalable tool for the Construction, Aggregation, Management, Exploration and Organization of healthcare data repositories. We will also deploy this system at the University of Miami Miller School of Medicine and demonstrate its utility within an active clinical research environment.

Cameo is focused on the acquisition, aggregation, security and facilitation of the exchange of healthcare information generated by point of care, and other transactional, medical informatics systems (e.g., hospital EMRs, analytical laboratories, clinical trials management systems). The system will enable healthcare data providers to make this primary data available for clinical research, such that it is secure, searchable, and anonymized. Since Cameo makes no attempt to re-represent data from the primary source systems and also leverages the existing capabilities of these systems it presents a relatively low barrier to entry for data providers who wish to make data available through the system.

Cameo is also complementary to efforts like caBig and can be made compatible with caGrid. Additionally, Cameo can enable organizations or vendors who would like to achieve caGrid compliance do so. By acting as a common point of integration, Cameo can help organizations who have achieved compliance with HL7 standards, but do not have sufficient resources to make their systems compatible with caGrid, achieve caGrid compliance.

Once data has been secured and aggregated by Cameo, it can also be made available for processing using any number of computational techniques, for data mining and analysis, or transformed into topic maps[vii] and other representations. These derived representations can then also be made available through Cameo. The initial implementation of Cameo will include the ability, on the part of data providers, to define relationships between the datasets under their control, and then create composite datasets based on these relationships.

The demonstration system will focus on the construction of repositories from HL7 messages, but is architected such that it can be easily extended to process healthcare data conforming to other standards (e.g., CDISC), or even data represented in proprietary or institutionally unique formats. The system is architected around a highly scalable service-oriented architecture and as a result, given sufficient resources, could be used to form the basis of a network capable of aggregating healthcare data from institutions throughout the United States.

The system also addresses the issue of trust between data providers and clinical investigators. Clinical research projects can be thwarted due to issues related to securing and anonymizing data[viii]. Cameo provides administration and security features for the data under its management. These security features allow data providers to be assured that the confidentiality requirements (e.g., HIPAA) of the information for which they are responsible are guaranteed. These security features also enable the system to act as a platform for collaboration.

Finally, Cameo provides tools for the rapid filtering and download of data under its control. Cameo's service-based architecture will also enable the system to be fairly easily extended to function as a secure ETL platform.

The Approach

We propose to build a system, which will act as a tool for the creation and management of healthcare data repositories. The proposed system will ensure stringent protection of individual privacy, while enabling use of aggregate healthcare data in clinical research. Additionally, the system will act as a channel to facilitate collaboration among clinical researchers. The initial system will aggregate data from any data source capable of producing HL7 messages. The system will provide tools for investigators to easily search and explore the repository, as well as tools to allow data providers to easily manage access to datasets.

We will also conduct a demonstration project with the system at the University of Miami Miller School of medicine. During this phase of the project, the system will be used to acquire data from a number of UM institutes and organizations, and also tested in the context of real clinical research applications.

Cameo is web-based system built around a service-oriented processing tier. Cameo is designed to allow organizations, that are originating sources of healthcare data (e.g., hospitals), to make this data available to other organizations and individuals who would like to use this data for clinical research. For the remainder of this proposal the organizations who are providing data will be referred to as providers and the individuals and organizations who are receiving this data will be referred to as investigators. Cameo allows providers to easily make their data available as aggregate sets that are searchable and downloadable by investigators. The system also provides services for security and anonymization, as well as tools for managing authorized access to these datasets.

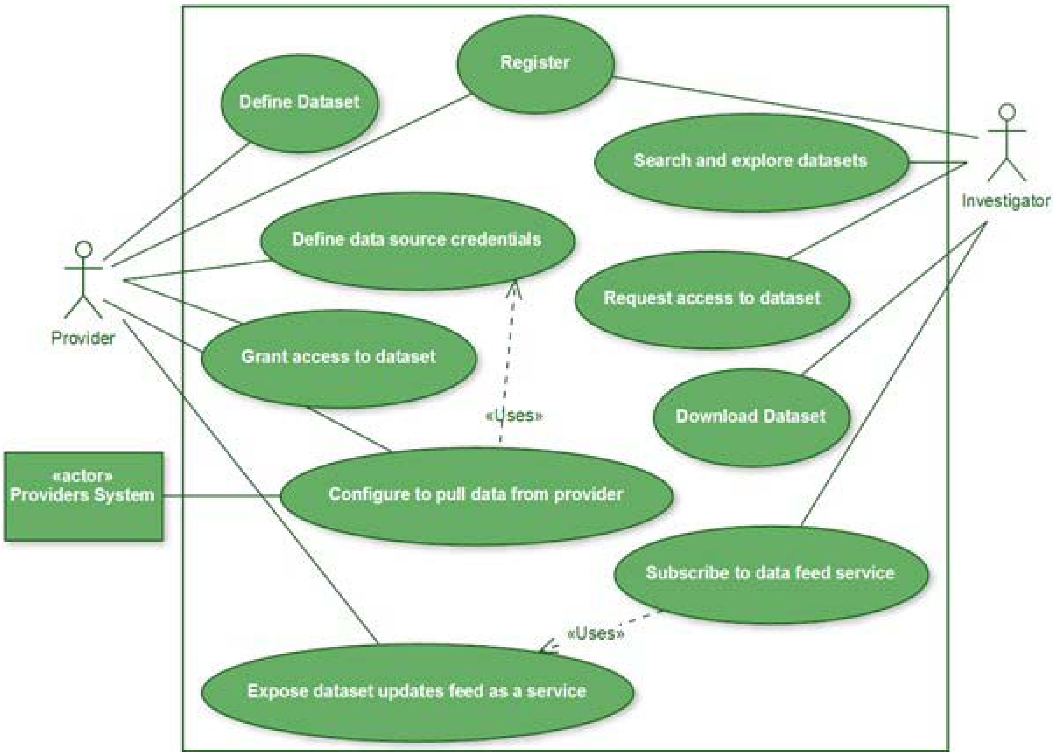
Investigators looking for data, with which to conduct clinical research, may use Cameo to conduct free-text searches in order to find datasets that meet their needs. Investigators may then use additional tools to explore summary statistics related to these datasets. If any available dataset is interesting to the investigator, then the investigator can use Cameo to request access to this dataset from the provider. Once granted access, the investigator may download copies or subsets of the data, and optionally subscribe to a feed (a webservice) in order to receive updates to the dataset. Investigators may download data for as long as they have authorized access.

All data managed by Cameo, in the case of the demonstration system, will conform to HL7 standards for representation (both v2 and v3). Providers will send data to the system using HL7 message streams, and investigators will access data as aggregate sets of HL7 messages or, for incremental updates, also as HL7 message streams. The system will make no attempt to re-represent this data in a unified data model, for mining or other uses. Cameo is designed to allow providers to easily publish and manage data, while enabling researchers to easily find and rapidly access this data. Data mining, semantic network and other applications can be built on top of, or using data provided through Cameo. Cameo is focused on developing a system and technology to meet the unmet need of providing a scalable mechanism for easily enabling the construction, and ensuring the security, of healthcare data repositories from across the spectrum of current sources of healthcare data.

Model use case

While the system will support other functionality, the primary use case Cameo is intended to support is that of providers aggregating, publishing, and authorizing use of primary healthcare data for secondary use in research. Under this model both providers and investigators will register with the system in order to gain access. First we describe the provider principle use case.

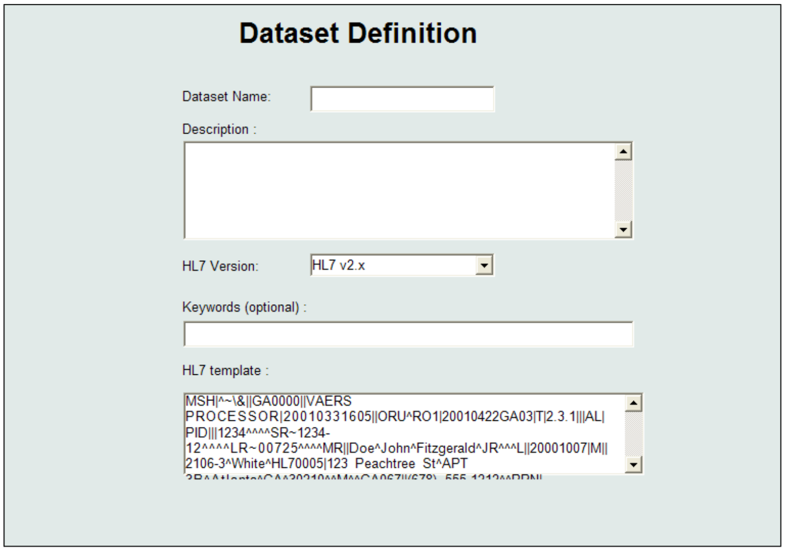
Figure 1 - Principle use cases


Register
Define Dataset
Search and explore datasets
Investigator
Define data source credentials
Provider
Request access to dataset
Grant access to dataset
«Uses»
Download Dataset
«actor»
Providers System
Configure to pull data from provider
Subscribe to data feed service
«Uses»
Expose dataset updates feed as a service

Providers who wish to use the system must first register. During the registration process the provider will supply sufficient information to identify the provider organization, as well as contact information for the administrator. For the initial demonstration project, all user accounts will be managed locally (by an overall administrator) in Cameo using OpenSSO. Extensions to Cameo, during the demonstration period, will provide the ability to delegate authentication to provider organizations, reducing risk to unauthorized access by giving the provider organization full control over access.

After registration providers may define and configure datasets to be managed using Cameo. Definition of a dataset primarily involves naming the dataset, providing a textual description of the dataset, and providing a template (or schema) for the structure of the data. In the case of the demonstration system, all templates (and therefore messages) must conform to either HL7 v2 or v3 standards. After the dataset is defined it must be configured and populated.

Figure 2 - Mockup of dataset definition



Dataset Definition

Dataset Name:

Description :

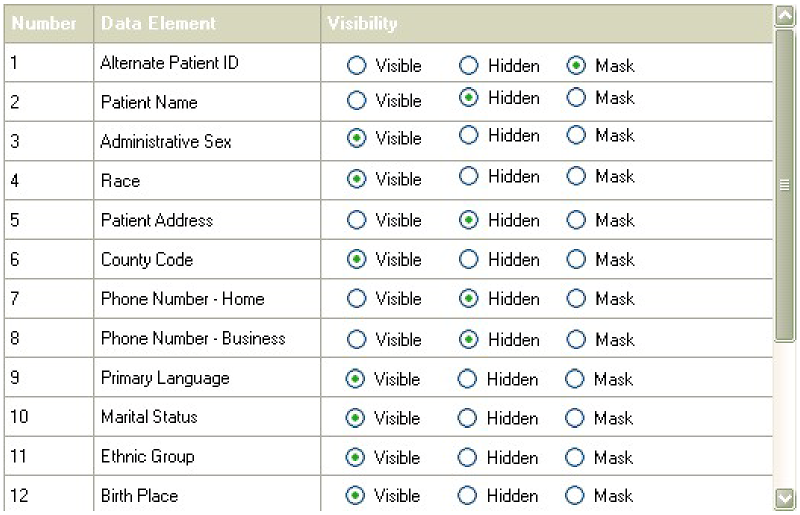
HL7 Version: HL7 v2.x

Keywords (optional) :

HL7 template :

```
MSH|^~\&||GA0000||VAERS
PROCESSOR|20010331605||ORU^RO1|20010422GA03|T|2.3.1|||AL|
PID|||1234^^^^SR~1234-
12^^^^LR~00725^^^^MR||Doe^John^Fitzgerald^JR^^^L||20001007|M||
2106-3^White^HL70005|123 Peachtree St^APT
3B^^Atlanta^GA^30210^^M^^GA067||(678) 555-1212^^PRN|
```

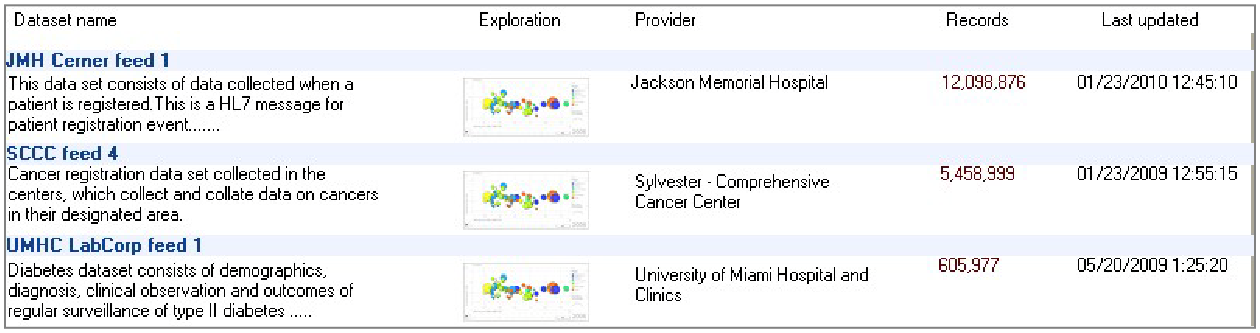
Datasets are configured by defining which fields in the template may be visible, hidden or masked. During dataset definition, a template for the dataset is provided. The system will parse this template, attempt to identify fields that could potentially contain personally identifying information (PII), and mark these to be hidden in the published dataset. During configuration the provider reviews these system suggested hidden field recommendations, accepts or declines them, selects additional fields to be hidden or selects fields to be masked. Fields selected for masking will have their values replaced by a system generated value (of the same datatype as the original) in the dataset that is made available to investigators. The system will maintain a secure dictionary of masked values mapped to original values.



Number	Data Element	Visibility		
1	Alternate Patient ID	○ Visible	○ Hidden	● Mask
2	Patient Name	○ Visible	● Hidden	○ Mask
3	Administrative Sex	● Visible	○ Hidden	○ Mask
4	Race	● Visible	○ Hidden	○ Mask
5	Patient Address	○ Visible	● Hidden	○ Mask
6	County Code	● Visible	○ Hidden	○ Mask
7	Phone Number - Home	○ Visible	● Hidden	○ Mask
8	Phone Number - Business	○ Visible	● Hidden	○ Mask
9	Primary Language	● Visible	○ Hidden	○ Mask
10	Marital Status	● Visible	○ Hidden	○ Mask
11	Ethnic Group	● Visible	○ Hidden	○ Mask
12	Birth Place	● Visible	○ Hidden	○ Mask

Figure 3 - Dataset field configuration

In some cases, data containing PII may be required in order to conduct useful research using information that spans datasets. Cameo will also support the definition of relationships (joins) between datasets. Providers supplying multiple distinct datasets may use tools to define relationships between these datasets based on fields, or common keys, present in each dataset. Once defined, Cameo can create a derived dataset from these relationships. The dataset can then be managed and made available in the same fashion as other datasets. After configuring visible, hidden and masked fields, providers must specify an HL7 feed (webservice) for the dataset. Specification of the feed requires providing a WSDL URL, SOAP or service endpoint URI and authentication parameters. Once configuration is complete the dataset can be populated. The primary mechanism of populating the dataset should be via the HL7 feed, which Cameo will begin pulling data from as soon as configuration is complete and the dataset has been activated. Alternately, a file containing a bulk set of data may be uploaded to initially populate the dataset.



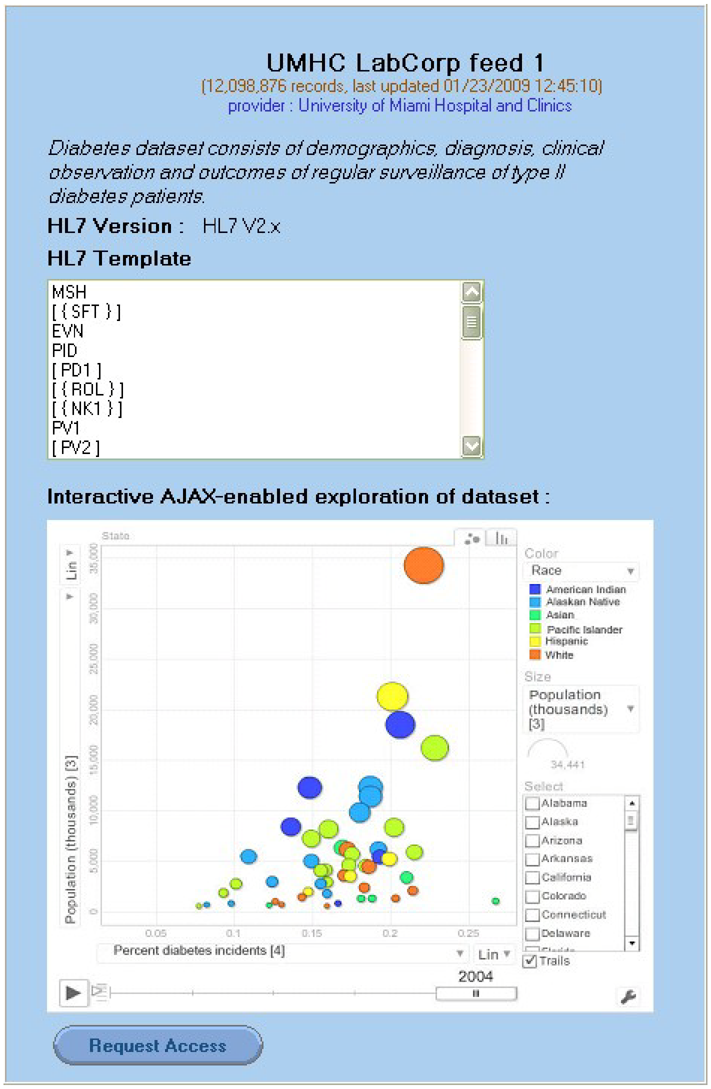
Dataset name	Exploration	Provider	Records	Last updated
JMH Cerner feed 1 This data set consists of data collected when a patient is registered.This is a HL7 message for patient registration event.......		Jackson Memorial Hospital	12,098,876	01/23/2010 12:45:10
SCCC feed 4 Cancer registration data set collected in the centers, which collect and collate data on cancers in their designated area.		Sylvester - Comprehensive Cancer Center	5,458,999	01/23/2009 12:55:15
UMHC LabCorp feed 1 Diabetes dataset consists of demographics, diagnosis, clinical observation and outcomes of regular surveillance of type II diabetes		University of Miami Hospital and Clinics	605,977	05/20/2009 1:25:20

Figure 4 - Search results

Access to datasets is granted based on requests from investigators. Providers use a management interface to review investigator requests for access (including a research plan), communicate with investigators, and grant or revoke access. The investigator usecase, described next, provides a description of how datasets are searched and accessed.

Investigators must also register to gain access to the system. After registration, investigators may use a simple text-based search tool to identify datasets of potential relevance to their research. The tool searches an index created from a combination of provider specific information (e.g., “institution name”), text descriptions of datasets (e.g., “diabetes”), any additional provided keywords, information contained in the HL7 template describing the dataset (e.g., marital status, race), as well as information contained in the dataset itself. Relevant hits are returned in a list containing summary information about the dataset.

Investigators may then drill-down to view details about a particular dataset. Dataset details include the HL7 template describing the dataset, summary statistics based on the distribution of values associated with the fields in the dataset, as well as an interactive



UMHC LabCorp feed 1

(12,098,876 records, last updated 01/23/2009 12:45:10)

provider : University of Miami Hospital and Clinics

Diabetes dataset consists of demographics, diagnosis, clinical observation and outcomes of regular surveillance of type II diabetes patients.

HL7 Version : HL7 V2.x

HL7 Template

```
MSH
[ { SFT } ]
EVN
PID
[ PD1 ]
[ { ROL } ]
[ { NK1 } ]
PV1
[ PV2 ]
```

Interactive AJAX-enabled exploration of dataset :

Figure 5 - Dataset details

graphical tool for viewing and exploring these statistics.

Within the dataset details interface, the investigator can choose to request access to the specific dataset being viewed. To request access, the investigator must provide a research plan for the use of the data. The request is sent, within the system, to the provider managing the specific dataset. Once the provider approves access, the investigator may download the entire current dataset or an approved subset. Additionally, the investigator may also configure and subscribe to a webservice for updates.

Cameo distributes authority to grant access and the responsibility of controlling viewable content to the data providers. The system provides tools for managing access to datasets and restricting access to subsets of these same datasets.

Cameo's service-based architecture (discussed in more detail below) will also enable the system to rapidly filter datasets and make them available for download. Datasets in Cameo are stored as aggregate sets of HL7 messages. This format makes distribution of filtering and assembly of datasets simple. The aggregate sets can be chunked and sent to independent distributed processes for filtering before being streamed for download. This same architecture also allows the system to be extended, by adding additional services, to act as a platform for rapid transformation of data to other formats, as well as other types of processing.

Architecture

Cameo is divided into two principle tiers: A web-based application layer which will support user interaction with the system and a service-based message and data processing layer, for pulling, validating, scrubbing, aggregating, filtering and otherwise processing HL7 data from the source systems. These two components will be separated by a firewall.

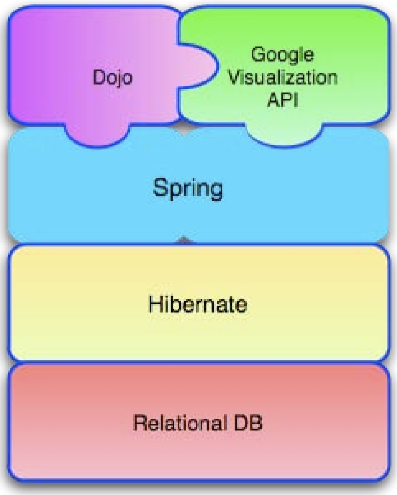


Figure 6 - End-user application stack

The application layer is used by end-users to define datasets, search datasets, download data, and perform all of the other end-user functionality described above. The end-user application will be built as a multi-tier web-based application, using both the Dojo toolkit and Google visualization API to enhance the presentation layer. It will use the Spring framework to implement the application logic layer. Hibernate will be used to enable data storage and access. The application will run against a platform neutral relational database data model (schema). This flexibility is largely enabled by using Hibernate to implement the object/relational mapping layer. For the demonstration system, we will use MySQL as the RDBMS. The data stored outside the firewall, in the application DB, will not contain any highly sensitive information (e.g., PII or authentication credentials). The application DB will only contain summary information about the datasets, indexes, and user preference and configuration information. Cameo follows standard best practices for data security. All sensitive information are isolated from unauthorized access behind multiple firewalls. Once data is sanctified and all PII is removed, Cameo will use a push mechanism across an encrypted channel to make data available for download by investigators. The communication channel will be verified by MAC verification and route data. Authentication and load permissions will be limited to a single non-privileged system account that is not available to any user. Once non-PII data is on the application server, access will be granted across the external firewall utilizing a standard AAA method described in the security section below. The entire Cameo system will be protected by several hardware firewalls utilizing Statefull Packet Inspection (SPI) and Intrusion Detection/Protection Systems (IDS/IPS).

The message and data processing layer deals with HL7 messages from data providers. This layer processes these messages, assembles and filters datasets, makes these datasets searchable and otherwise enables processing of the HL7 data. Once a dataset has been defined by a provider, including appropriate webservice connection and authentication information, Cameo will begin pulling data from the defined webservice. Key services in the processing pipeline are described here.

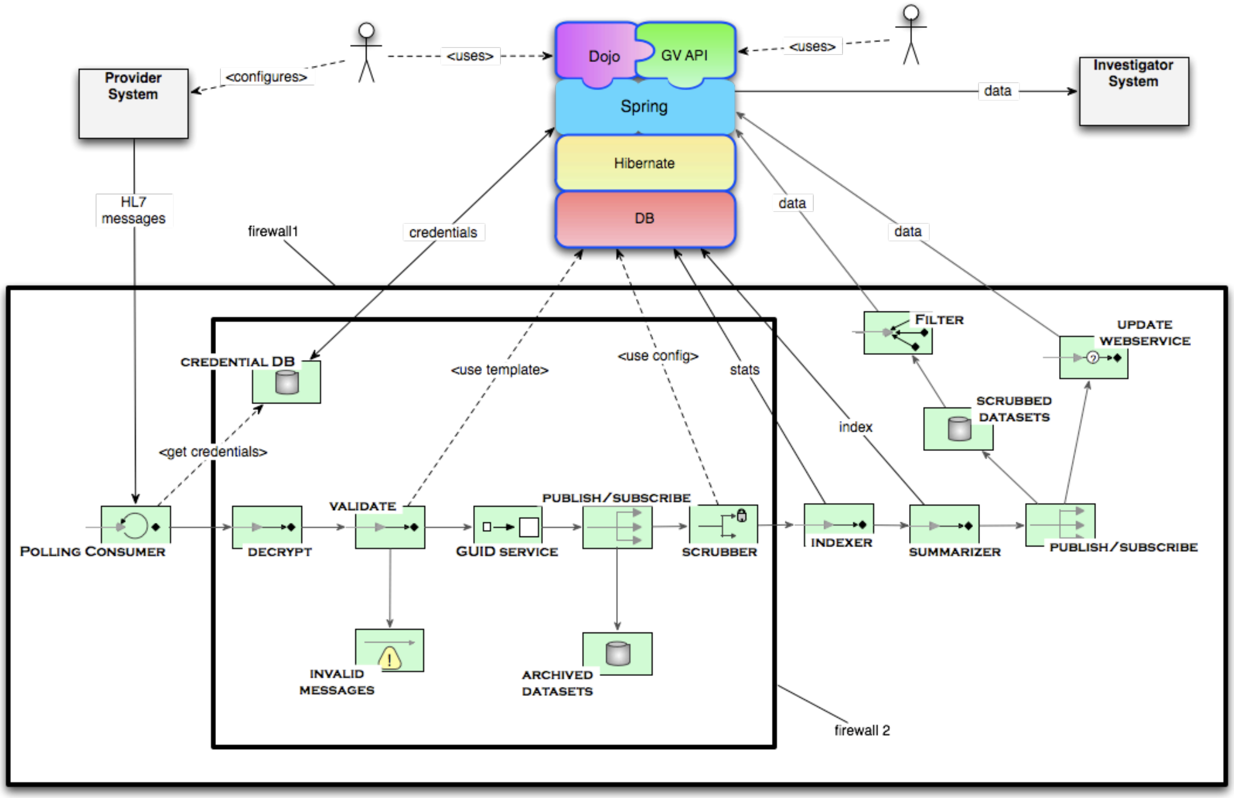


Figure 7 - Processing tier key components

Decryption service: All in-transit messages received from the provider should be encrypted to protect against unauthorized access or modification. The decryption service provides functionality for handling encrypted (e.g., triple DES) messages.

Validation service: After decryption, messages must be validated against the HL7 template provided when the dataset was defined. The validation service provides validation processing, and in the event that a message fails validation, routes this message to an invalid message queue for error handling. The administrator of the dataset will be notified of the error. Invalid messages will be held for a specified amount of time before being purged.

GUID service: All messages received by the system will given a globally unique identifier and routed to a secure storage environment. This secure store will contain copies of all valid messages received by the system. These messages can then be made available for future processing, such as data or text mining applications.

Scrubber: After being tagged with a unique identifier, a copy of the complete message is also routed to a service that will scrub the message of all PII. The scrubber will use the configuration information entered when the dataset was defined.

Summarizer: The summarization service will compute and update various summary statistics based on the attributes of the dataset associated with the message. These statistics are then made available to the end-users through the web-based UI.

I**ndexer:** This service indexes the message to make its content searchable using free-text search by investigators. At least two levels of indexing will be implemented in the demonstration system. One level of indexing will be available to all investigators across all datasets. This level will provide search capabilities based on message content, but will only return summary information about the dataset associated with any relevant messages. Once an investigator has been granted access to a dataset, then search results will include the option of being to review the the full set of messages which satisfy the investigator's search.

After being indexed, the scrubbed message is combined in the complete dataset. Optionally, it may also be routed to a publish/subscribe queue in the case where investigators have configured a webservice from which to receive updates.

Aside from the basic message processing pipeline, other services will be constructed to enable rapid filtering and downloading of datasets, as well as the assembly of derived datasets based on provider defined relationships between datasets.

Development strategy
Cameo will be a predominately Java-based system. To build the above described system, we will establish an engineering team consisting of a senior Java software engineer and two dedicated junior Java software engineers. This team will work together using established software engineering practices, including following documented coding standards and conducting code reviews. The group will have a variety of software engineering tools at its disposal, including Subversion[ix] for source code control, Eclipse[x] for software development, and Caretta[xi] for user interface design and simulation.

Specific Aim1 – Develop a scalable software system for secure aggregation and distribution of healthcare data from primary source systems

Construction of the message and data processing tier directly addresses the goal set in Aim 1. Building this component enables secure aggregation, anonymization, and distribution of healthcare data. Construction of this component also enables the development of the UI described in Aim 2. The message and data processing tier will be built on a service-oriented architecture using Mule[xii] as the backbone messaging framework. We will also be using the Mirth[xiii] HL7 messaging framework to implement a number of the key components needed within the processing layer. Mirth has been built on top of Mule and provides numerous utilities for handling both HL7 v2 and v3 messages. Both Mule and Mirth are proven open source systems with active developer communities and numerous deployments.

Indexing of datasets and other information will be done using a service built on the Apache Lucene[xiv] text search engine library. We may also conduct research projects using both the Minion[xv] and Xapian[xvi] libraries.

Specific Aim2 – Develop a user interface for the management, search, exploration and download of data contained within the repository described in Aim1

Construction of the end-user application, with the functionality described above, satisfies Aim 2. The end-user application will be implemented using Java[xvii], Spring[xviii] and Hibernate[xix]. The user interface will be built using Java, the Dojo toolkit[xx], the Google visualization API[xxi].

All libraries and component systems used to build Cameo will be either developed internally, or available as open source systems. At the conclusion of the project, all code, documentation and other materials associated with Cameo will be made available as an open source offering.

Specific Aim3 – Deploy the system at the University of Miami Miller School of Medicine and demonstrate its utility within an active clinical research environment

To meet Aim 3 we will need to address issues of project management and communication in addition to technical issues. We will need to recruit collaborators to use the system and support these collaborators in their use of the system.

We have already identified collaborators, who have also provided letters of support, who are interested in using the system once it has been deployed. During the first year of the project we will also be actively recruiting other collaborators for the demonstration phase. Finally, we have also assembled a team that includes a physician/scientist, and project manager who will be facilitating interaction with our collaborators during the demonstration phase, and who will ensure that the system is indeed meeting our objectives from the point of view of its usability and value to clinical researchers.

Also, for the actual deployment, in addition to the hardware budgeted for the project, we will also have available all of the resources from the Center for Computational Science (CCS) described in the "*Resources*" section below.

Security
Healthcare information messages hold facts about personal and confidential information. When transporting healthcare information, the security and privacy of the information must be insured. To address these security needs the following common protections and security requirements need to be fulfilled:

Identification / Authentication: OpenSSO[xxii] will be the identity management and authentication service for Cameo. All clients for Cameo will be provided a user ID and strong password. The user will present this user ID and password whenever sending transactions. Single sign on will be implemented, which will enable users to authenticate once and gain access to multiple features. Wherever possible the authentication will be delegated to the user (provider/investigator) by authenticating the users against their organizations authentication service through OpenSSO. This delegation will mitigate the risk of unauthorized access to data as the user or user's organization is responsible for granting access. However, there will be facility for creating local users for Cameo in order to enable access to the users who do not approve integration to their organizational authentication service.

Authorization / Access control: OpenSSO along with the JAAS(Java authentication and authorization service) API will be used to provide access to system resources. Policies will be defined for access control of resources. Role based access control will be implemented. Each dataset will have a access control list and the provider will be the owner of the ACL and has complete control of providing access to the investigators for that dataset.

Integrity and confidentiality: It is important to ensure data is neither altered or available for unauthorized access. Encryption will be used to safeguard the confidentiality of the messages. All incoming and outgoing messages are Triple DES encrypted and logged.

Secured Channel: Secure communication channels (HTTPS) will be used to safeguard messages and data. When using userID/password and any other confidential data, application programs will issue a HTTP POST transaction. HTTPS will be used as the transport layer protocol for the transactions. This will ensure that the message is encrypted throughout the communication channel.

Audit: Audit trial and logs will be maintained for all the events in Cameo. This will ensure Non-repudiation to a certain extent , as accountability for all actions can be tracked. This will satisfy all attestation needs.

Privacy protection: Anonymization and Pseudonymization concepts will be used for privacy protection. Personally identifying information (PII) will be removed from the HL7 message before use in secondary contexts is allowed. In cases where re-identification of PII is required, Pseudonymization will be used. This will be done by identifying the list of possible PII fields and presenting it to the provider for identifying the fields to be anonymized or pseudonymized. HL7 Anon utility is a current example a utility for removing PII in HL7 messages.

Resources available at the University of Miami

The work done under this proposal will have access to current computational and storage resources available within the Center for Computational Science (CCS). The Center's facilities include over four Teraflops of computational power split between two distinct architectural paradigms, Symmetric Multiprocessing (SMP) and Massively/Embarrassingly Parallel Processing (M/EPP). More specifically, the UM CCS currently has a 256 CPU IBM Power5+ SMP cluster with high speed interconnect, a 96 CPU IBM Power4 SMP cluster and a 32 CPU SUN SPARCIIIi server. Additionally the Center has a Linux Xeon cluster with 256 processor cores for M/EPP programs and algorithms. This cluster has an aggregate of 512 Gigabytes of RAM. The cluster runs RedHat Enterprise Linux. In aggregate the Center currently has over compute CPUs with 1 TB of memory. There are over 100TB of shared SAN storage with a common namespace for all servers. The storage design is centered around IBM's GPFS. GPFS allows high speed access to the clustered storage through a distributed meta-store/object-store scheme. The presentation layer of the storage is done through Samba, which affords all clients equal access to the storage, utilizing a single management point. The center also provides full disaster recovery facilities with archive capabilities.

Cameo will be given high priority access to these resources for all development and proof of concept work.

In addition to computational resources, CCS will provide access to the FUSE (Flexible User Storage Environment) for this grant. FUSE represents abstracted storage access for both Relational and Non-relational data. Relational data services are provided through our MySQL and Oracle environments. These environments are highly scalable and resilient in design. Our relational storage provides an aggregate space in excess of 2 TB high performance relational storage. Non-relational storage will be provisioned using a combination of high performance parallel access for computation and deep storage pools for non-transactional data. These data can be accessed through all standard protocols for client access. Finally, Cameo will also have access to a secure pool of storage for all PII/Sensitive data. This pool will be isolated from the CCS and UM environment both physically and logically. Audit trails and access attestation will be provided for this storage pool.

Timeline and Milestones
The project will have two main phases. During the initial phase we will design and build the initial version of Cameo. This initial phase of the project will last approximately 64 weeks. We will also recruit collaborators, within the University of Miami Miller School of Medicine, during the first year of the project who will be providing data and/or using the system during the demonstration phase. The second phase will last the remainder of the second year of funding. During this phase we will operate and further test the system at the Miller School of Medicine. We will also prepare documentation for all code, as well as the code itself, for release as an open source offering. The project milestones, broken down by year, are as follows:

- Complete System Design (T+16 weeks)
- Complete Build 1 of message processing tier (T+31 weeks)
- Complete selection of initial collaborators for demonstration phase (T+52 weeks)
- Complete Build 1 of end-user application (T+55 weeks) – AIM2
- Complete Build 1 of full system (T+59 weeks) – AIM1
- Demonstration System ready (T+64 weeks) – AIM3
- Open source release (T+102 weeks)

The current timeline assumes that the message processing tier and end-user tier will not be built in parallel, which is the least aggressive method for estimating the time to complete construction of the system.

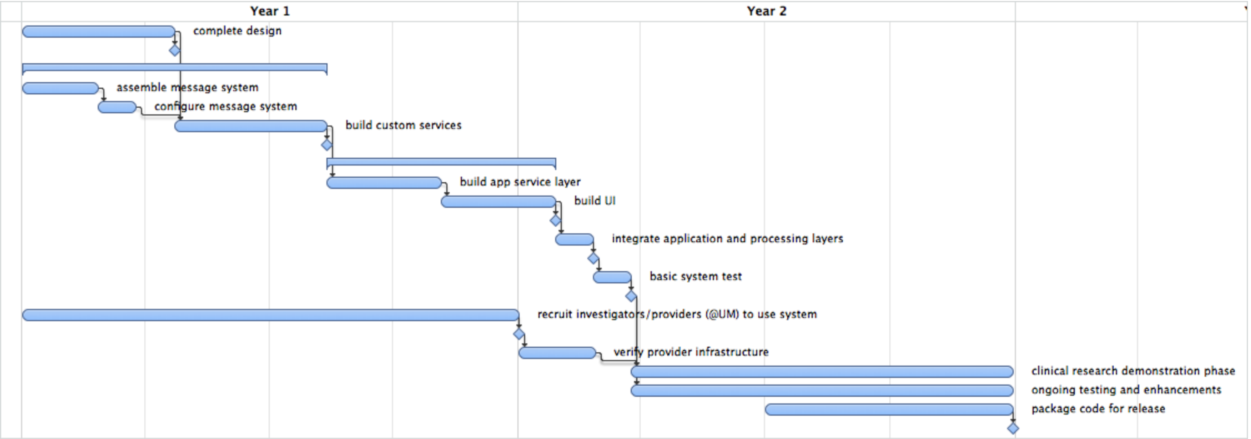


Figure 8 - Project timeline