The CTSA Ontology Mapper and Discovery Suite:
A Rules-Based Approach to Integrated Data Repository Deployment

A proposal in response to RFP-08-001

Technical Proposal

Primary Contacts:

Russell J. Cucina, MD, MS

Assistant Health Sciences Professor of Medicine and Center for Clinical and Translational Informatics

UCSF Clinical and Translational Science Institute

Associate Medical Director of Information Technology, UCSF Medical Center

University of California San Francisco

Robert Wynden

Project Director and Systems Architect – IDR Project

Academic Research Systems: OAAIS

University of California San Francisco

Phone: (415) 439-9756

rcucina@medicine.ucsf.edu

rob.wynden@ucsf.edu

TABLE OF CONTENTS

1. INTRODUCTION

The traditional approach to data warehouse construction is to heavily reorganize and frequently to modify source data in an attempt to represent that information within a single database schema. This information technology perspective on data warehouse design is not well suited for the construction of data warehouses to support translational biomedical science.

The Ontology Mapping approach presented below represents an alternative to a forced common agreement on a generalized representation of all source data in translational biomedical science. That is not to say that a common data model is not possible and preferred at some institutions with different system requirements. A generalized schema approach has inherent strengths; including leveraging well documented commercial IT processes to produce a production environment faster. Additionally, blended architectures which encompass the basic elements of both approaches may be possible. However, the challenges discussed here suggest an alternative approach is required.

Challenges Posed by the Traditional Approach

A data warehouse that combines clinical, clinical research, biomedical, biosciences, economic, administrative, and public health data to support translational research will hereinafter be referred to as an Integrated Data Repository, or IDR. There are several challenges posed by Integrated Data Repository projects which do not apply to the construction of most commercial warehouse implementations.

1. Integrity of Source Data

A clear requirement in the construction of an IDR is that source data may never be altered, nor may their interpretation be altered. Researchers will at times require clear visibility to the source data in its native format to verify it has not been altered.

2. High Variability in Source Schema Designs

IDRs import data from a very large set of unique software environments, from multiple institutions, each with its own unique schema.

3. Limited Resources for the Data Governance of Standardization

Widespread agreement on the interpretation, mapping and standardization of source data that has been encoded using many different ontologies over a long period of time may be infeasible. In some cases the owners of the data may not even be available to work on data standardization projects, particularly in the case of historical data.

4. Limited Availability of Software Engineering Staff with Specialized Skill Sets

Modification of source data during the data import process requires a large and highly skilled technical staff with domain expertise, talent often not available or only at considerable expense.

5. Valid yet Contradictory Representations of Data

There are valid, yet contradictory interpretations of source data depending on the domain of discourse of the researcher. Examples related to the inconsistency of the researchers' domain of discourse include:

- Two organizations may interpret the same privacy code differently.
- Researchers within the same specialty may not use the same ontology
- Clinical and Research databases often encode race and ethnicity in differing ways

The RFP states: “While significant resources do exist for some types of data transfer or sharing, the ability to integrate the various sources of clinical, laboratory, and genetic data is a significant problem for many individual or small research teams.”

We propose that this problem is a *primary* impediment for small and medium sized research teams.

We seek to facilitate creation of Integrated Data Repositories (IDRs) across CTSAs by directly addressing the urgent need for terminology and ontology mapping in translational science. An ontology mapping component is essential for providing successful and cost effective data integration for small- to medium-sized translational studies through seven complementary aims:

Theme 1 – Streamline data acquisition and identification process

Aim 1: Deliver data to researchers in a just-in-time fashion, instead of requiring that all data be transmitted to the IDR via a single common format and without the requirement that all data be stored within a single centralized database schema. This just-in-time approach to data translation will alleviate the barrier to adoption caused by the high project cost and the long project lifecycle associated with traditional data governance.

Aim 2: Provide a data discovery and data request user interface that allows researcher’s data requests to be semi-automated. With appropriate permissions from an IRB, researchers will be able to submit requests for data online, track the fulfillment of those requests by business analysts, and receive data sets within a secure environment. A reusable and collaborative self-service infrastructure for CTSA IDR implementations will drive down the costs associated with IDR system deployments and promote investigator uptake.

Aim 3: Facilitate the emergence of a commercial ontology mapping service market through the enhancement of the HL7 CTS II protocol to include standard electronic communications between CTSA sites and for-profit ontology mapping services. These commercial entities would then offer mapping services for large, established ontologies and for datasets derived from large and established software environments.

Theme 2 – Develop standards-based technical infrastructure

Aim 4: Provide the software infrastructure with which an IDR can provide data to researchers, organized by a hierarchical terminology appropriate to that researcher’s domain of expertise. Data can be organized simultaneously into multiple data hierarchies and terminologies. This will allow clinical and translational researchers to utilize the same data sets for various purposes irrespective of specialty or the type of study.

Aim 5: Provide a reusable and fully functional software component that can be installed at any CTSA site. The CTSA ontology mapper will be developed as an open source server environment. Full documentation describing the system design and installation procedure will be provided. An ontology mapper user group will be created to facilitate end-user adoption and to gather feedback from production sites.

Aim 6: Leverage the Protégé knowledge management system and the Prompt ontology mapping tool (both from the NIH-funded National Center for Biomedical Ontology) to provide advanced, yet intuitive, tools to enable less technical users to translate the complex array of data fields needed to fulfill data requests. If IDRs require Programmer Analysts or Database Analysts to fill requests, the cost of implementation would increase significantly and become an implementation bottleneck.

Aim 7: Facilitate inter-institutional data sharing by translating data definitions among one or more site-specific terminologies or ontologies, and shareable aggregated data sets.

2. TECHNICAL APPROACH

2.a. Inference Based Ontology Mapping

We propose an ontology mapping software service that runs inside of the Integrated Data Repository. This service will provide the capability to map data encoded with different ontologies into a format appropriate for a single area of specialty, without preempting further mapping of that same data for other purposes. This approach would represent a fundamental shift in both the representation of data within the IDR and a shift in how resources are allocated for servicing Translational Biomedical Informatics environments. Instead of relying on an inflexible, pre-specified data governance and data model, the proposed architecture instead shifts resources to handling user requests for data access via dynamically constructed views of data. Data interpretation happens as a result of a specific researcher request, and therefore only as it is deemed useful.

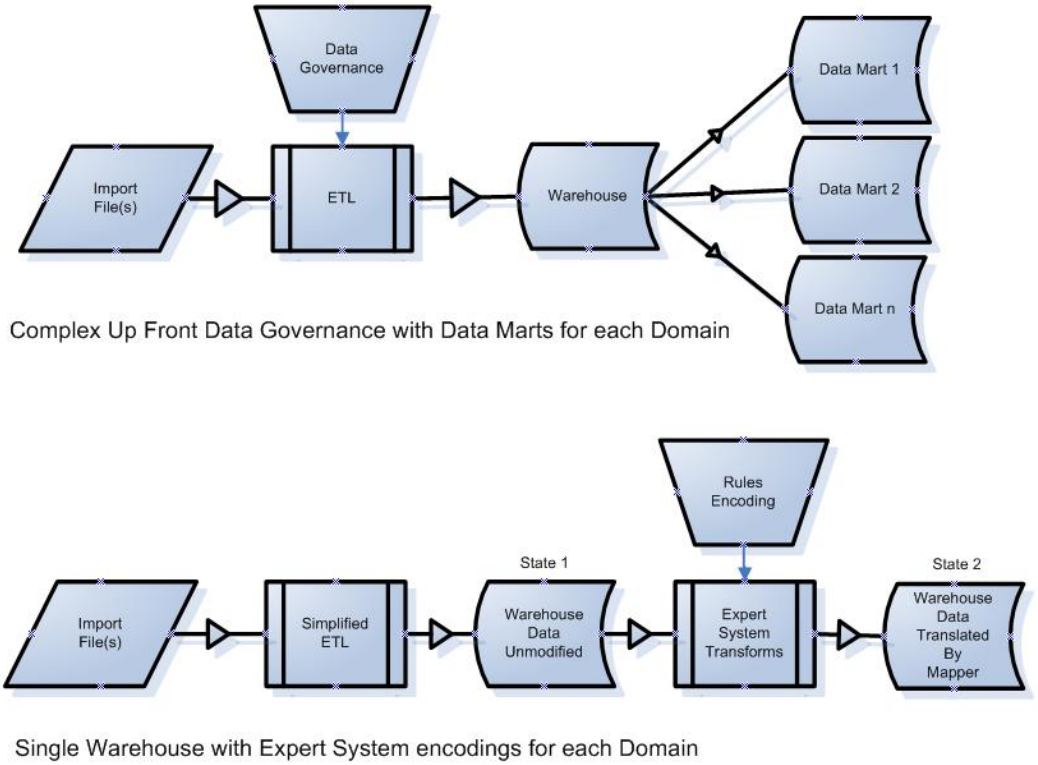


Figure 1. *Complex data governance (top) can be exchanged for rules encoding (bottom).*

The data in the bottom diagram will exist in both State 1 and State 2 simultaneously. The source data is always maintained within the warehouse in its source format as well as in its translated formats. Data may exist only once in State 1, but that same data may be translated multiple times and therefore may exist in several different formats in State 2. This is consistent with Theme 2 above, which aims to organize the same data using differing hierarchies and terminology encodings.

2.b. Resource Availability and Expense

Throughout this design we maintain a unified theme: the proposed methods allow an IDR to trade expensive up-front data governance for less expensive ontology mapping. Up-front data governance employs the time of expensive and highly trained staff with backgrounds for determining acceptable governance processes and standards to reach consensus from major stake holders. Up-front governance also requires highly technical expertise for control of ETL (Extract, Transform and Load) processes. This design aims to lower the capital

investment required to implement an IDR, by instead favoring personnel with a business analyst or equivalent background.

2.c. Interaction Model

User interaction with an IDR that implements the proposed tools will differ from that of a traditional data warehouse in two important respects:

Discovery: In models where up-front data governance has been applied, the data governance and standardization process will generate a large amount of documentation required to describe the source data, raising a barrier to researcher utilization. In the proposed model, the knowledge required of the researcher will be significantly reduced, and the researcher would only require enough information about the data available to formulate a specific request for access.

Translation: The translation of data from its source ontology into the ontology required by the researcher will not be completed during ETL. The ontology mapping will be completed after the source data has already been imported into the IDR.

To support these distinctions, we will develop two technologies that will make this approach practical:

1. Discovery Interface – Since all source data will not be analyzed in detail at the time of the initial ETL process that brings data in to the warehouse, a mechanism is required to conceptualize the IDR contents. A web browser-based interface for data discovery and concept mapping will be used to describe the contents of the IDR so that the researcher can learn what types of data are available prior to requesting IRB approval for access. This self-service user interface is described below.

2. Inference Based Ontology Mapping – The source data must be translated into the ontology that the researcher requires for a particular domain of expertise. The IDR will use a rules-based system to perform this mapping of source data format to the researcher's ontology of choice.

2.d. Discovery Interface

IDR's that adopt these tools will utilize the Discovery Interface for browsing and requesting access to data. The Discovery Interface will be published into web pages that can be displayed within an enterprise web portal environment, but will be accessible by any user equipped with a modern web browser, regardless of their operating system of choice. Researchers will be granted role-based access to the Discovery Interface (but not to any source data) prior to IRB approval.

The Discovery Interface will provide the following specific features:

1. A full conceptual view of the data contained within the IDR that describes what the data is and the relationships among data.

2. The interface will adhere to familiar user interaction models to make the interface as easy to learn and use as possible. New or experimental data visualizations will be presented only as a last resort when necessary.

3. Interactive AJAX (Asynchronous Javascript and XML) based drop down menus for access to hyperlinks and to related system components.

4. The ability for the System Administrator to import new dataset descriptions into the IDR and to subsequently publish that information for access by researchers.

5. A description of the specific ontology used to encode each source datum.

6. Access to simulated data that show examples of the specific formatting for each particular conceptual element.

7. Access to concept maps that document the environment within which the data were collected. (These graphical maps will look similar to workflow maps generated in industry for business process automation.) Concept map diagrams will be used to encode information about the context within which the data was generated. For example, what other business processes or automated systems were involved in the collection of the data. These graphical representations of business process will provide contextual information to the researcher visually and quickly.

8. Help text providing a written description of each particular conceptual element

9. Access to the name of the source data environment from which the conceptual element was imported

10. Access to the date of the last successful data import and next expected import date as well as other related metadata (name of the system owner, contact info etc.)

11. Access to researcher annotations regarding each specific conceptual element using a web based annotation interface.

12. If pertinent and available, a link will be provided to the source data owner's website.

Below are conceptual screen shots of the proposed Data Selection interface:

Data Selection Tab

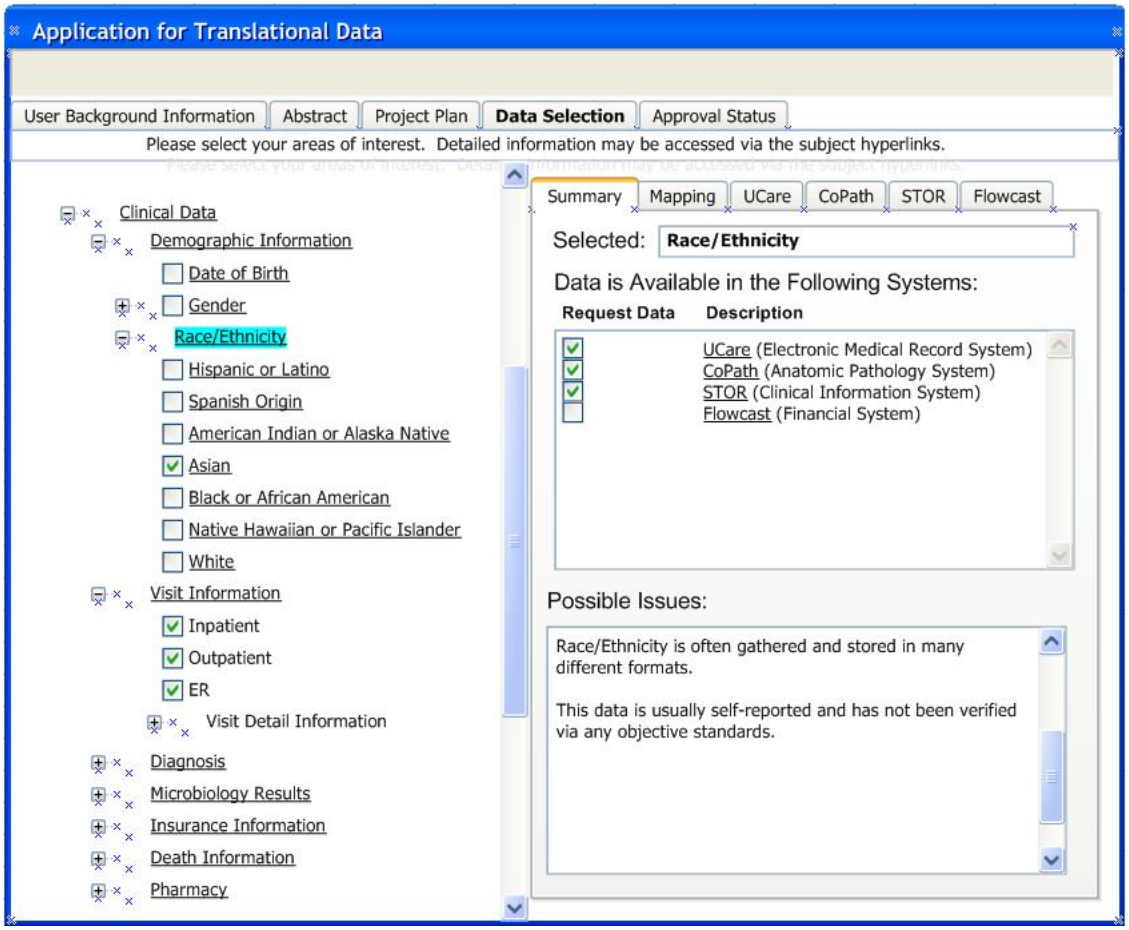


Figure 2. *The Data Selection interface portion of the Data Discovery Interface.*

This Data Selection Screen has the following main components:

- A scrollable list of data elements. This list describes data element types in general terms. The user can click on an item shown here to request access to information of this type.
- A properties pane synchronized to the selected element in the data element list. The items described in the scrollable list are hyperlinks; clicking on one these causes the properties pane to refresh and resynchronize to detailed information regarding that specific item.
- The summary tab describes the datasets available that contain information of this type.
- For each dataset, the researcher may specify the source systems of interest.
- A description of common issues regarding information of this sort is supplied.

Mapping Tab

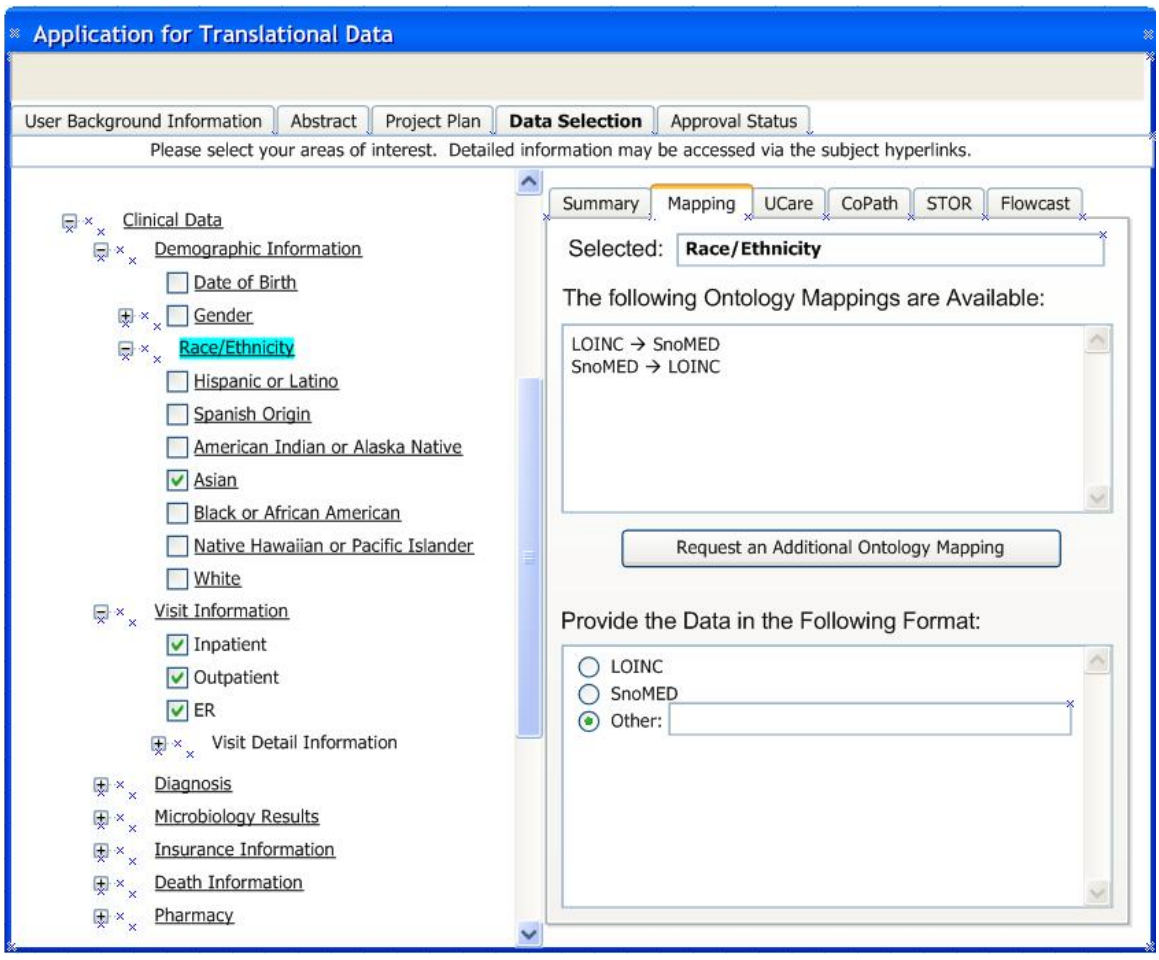


Figure 3. *The Mapping Selection interface portion of the Data Discovery Interface.*

This Mapping Selection Screen has the following main components:

- A scrollable list of the ontology mappings available for this data element type.
- The ability to request creation of a new ontology mapping rule.
- The ability to select a specific mapping for the data request.

Dataset Tabs

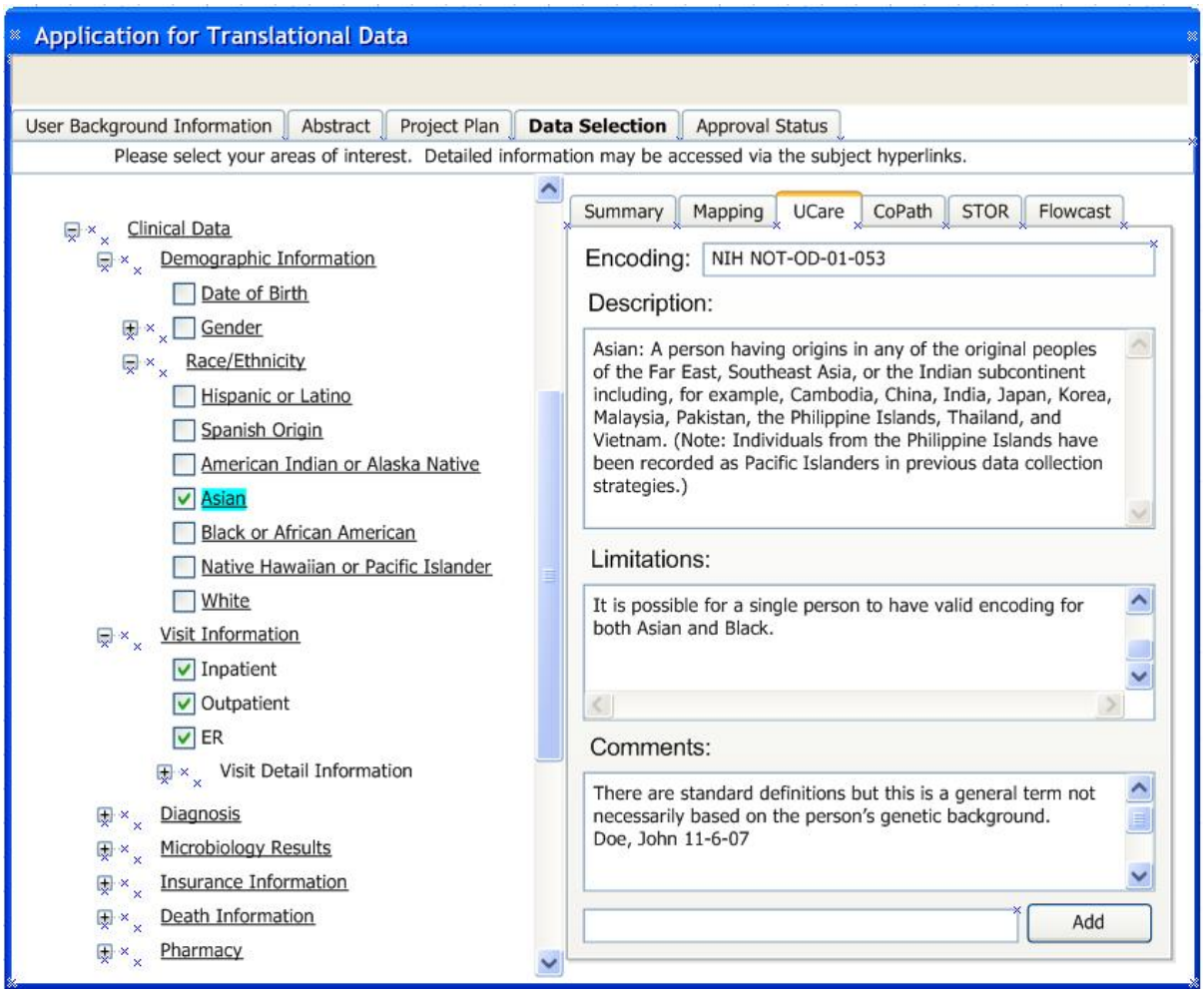


Figure 4. *An example Dataset tab within the Data Discovery Interface*

The Dataset Tabs have the following main components:

- The encoding used to represent that data element within this specific dataset.
- A written description of this data element.
- A written description of any known limitations of the data.
- An area where users may make their own comments that can later be viewed by other researchers.

The above data selection screens are at the heart of the Discovery Interface. However, researchers need to understand not only the nature of the data but the environment within which the data were originally captured and used. Descriptions of this sort are particularly difficult within a standardized textual format. Therefore, instead of favoring lengthy and complex written descriptions of the operating environments, the system will instead employ concept maps.

Concept Map Views

The concept map is an element of knowledge management theory first proposed by Joseph Novak in the 1970s [4]. Concept maps have since gained wide acceptance in the education and business process automation fields. These proposed tools will use concept maps as the primary mechanism of describing the operating environment within which the data were first gathered and used.

Within the dataset tab interface those datasets for which concept maps are available will display a list of icons representing each available concept map. By hovering over each icon a popup summary of the corresponding map will be displayed.

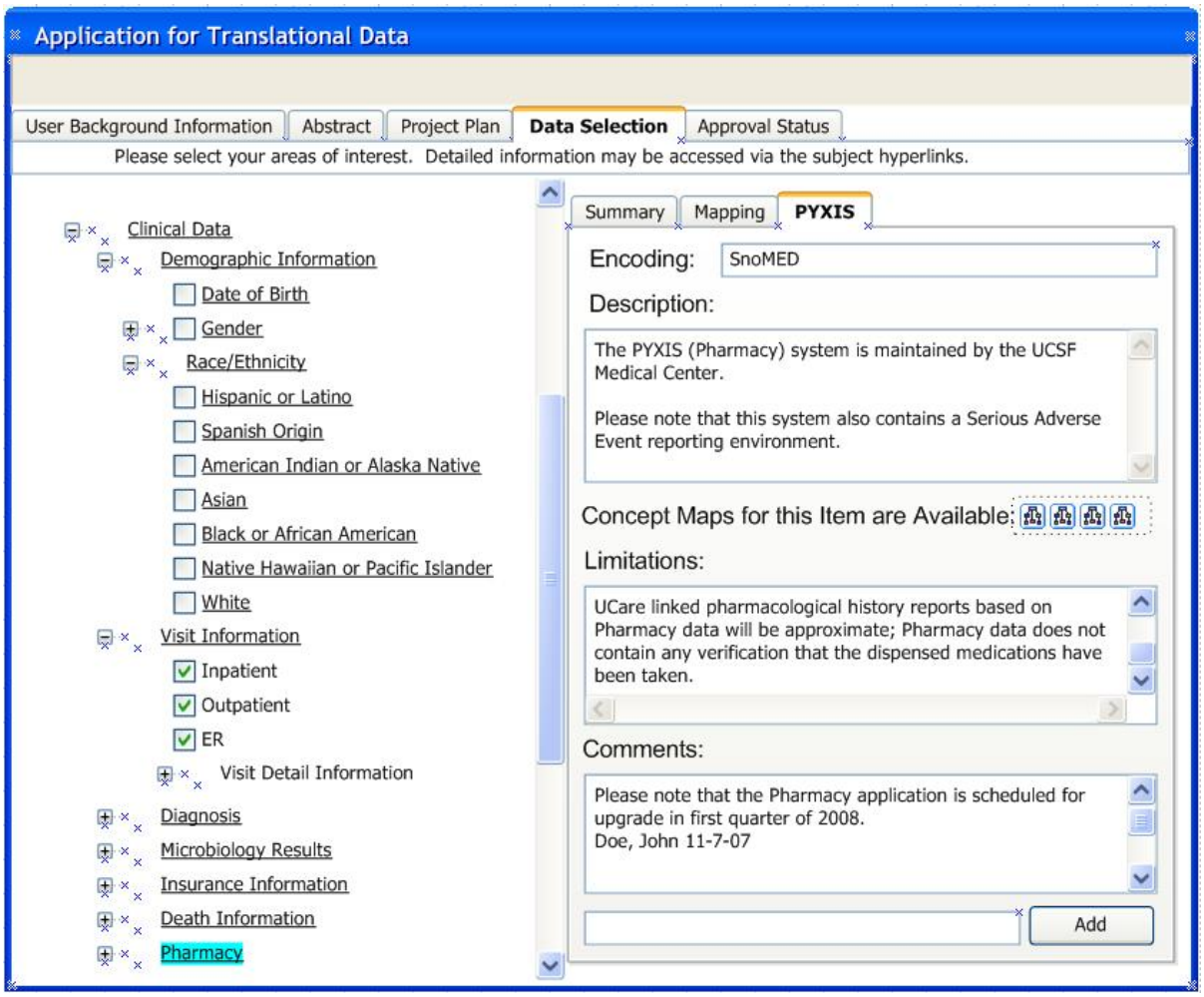


Figure 5.) A *Dataset Tab with available Concept Maps*.

If the researcher clicks on a particular concept map icon a popup window will be displayed describing the workflow of the environment where the data were collected.

The CMapTools system will be used as part of this design. CMapTools is a web browser-based open-source framework for visualizing and interacting with concept maps developed by the Institute for Human and Machine Cognition (IHMC). The proposed tool set will not contain features for authoring concept maps, nor will it specifically expose authoring environments for mapping rules. It will instead rely on existing commercial and open source environments for those tasks. To import files encoding concept maps and rules the system will support standard formats such as the RDF (Resource Description Framework) format. Once imported these concept maps will be housed within the IDR itself.

These concept maps will be drawn by the Business Analyst while servicing data requests. They will be used to document the software environments and business processes used to generate the source data. These graphical charts will offer the researcher a quick visual method of understanding the environment within which the data was generated and used.

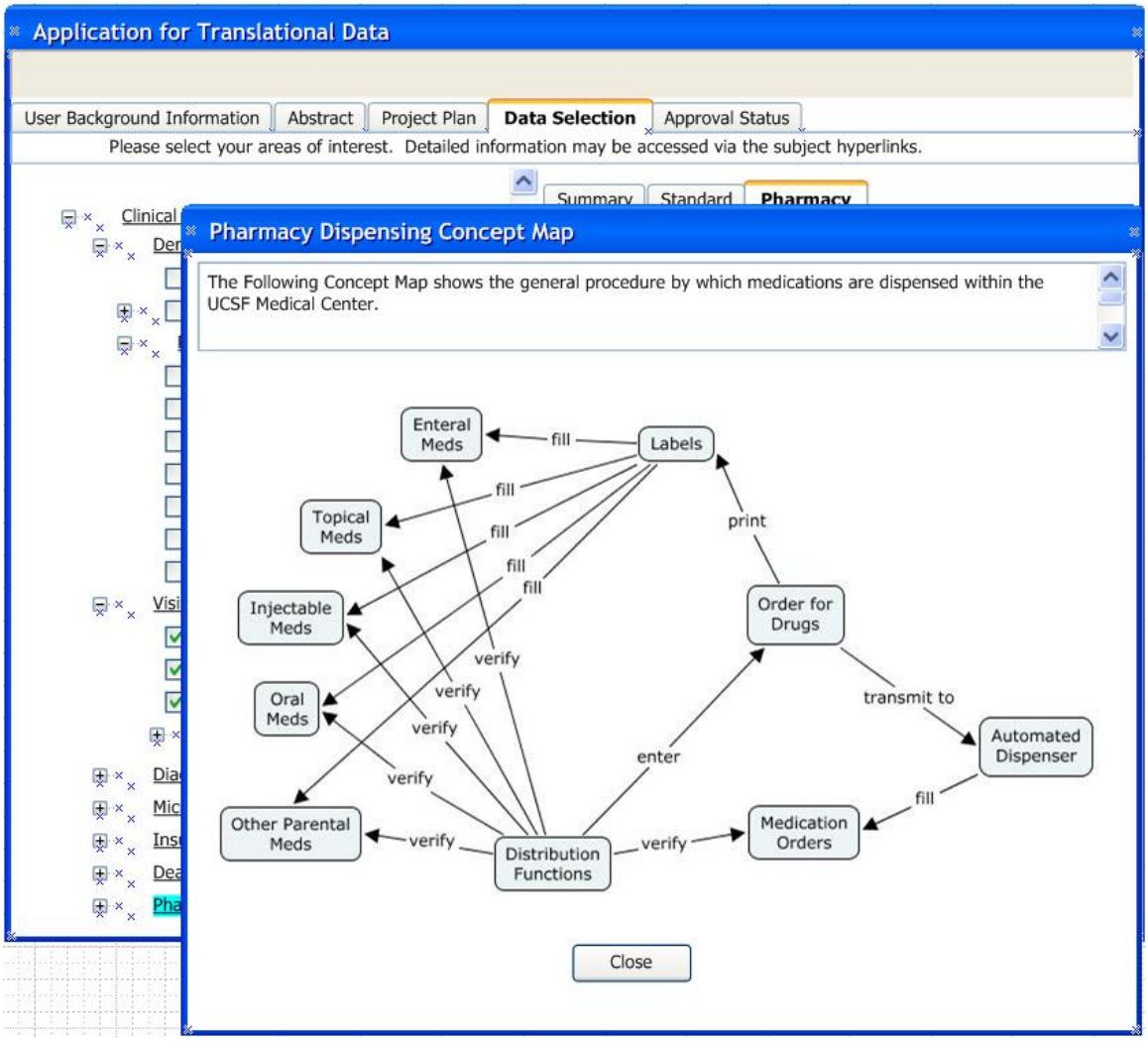


Figure 6. A demonstration concept map describing the operations of an operating environment within which data is generated for a particular class of data elements (pharmacy data).

2.e. Capturing User Requests

The Discovery Interface will capture new requests for user access. This work builds on previous implementations at the University of Rochester translational warehouse system.

The "Request New User" interface will provide the following features:

1. Prompt for user background information required by the IRB, including the name and contact information of the Principle Investigator, research staff, and administrative contact. This screen will completely describe exactly who will be granted access to the data.

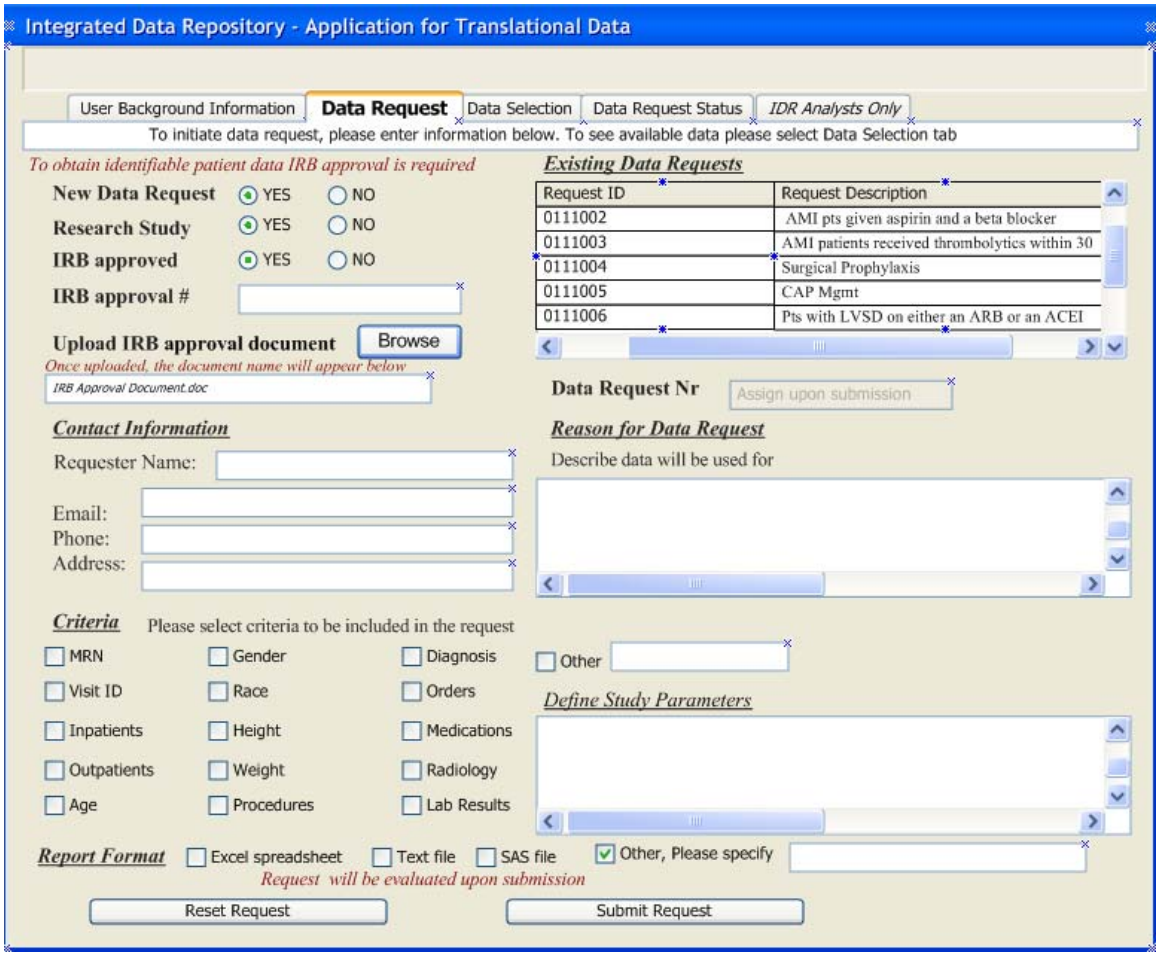
2. Prompt for the abstract that describes the research to be conducted, including relevant background information.

3. An outline of the proposed project plan.

4. A data selection screen that allows the researcher to browse what data are available in the system and select the data elements of interest.

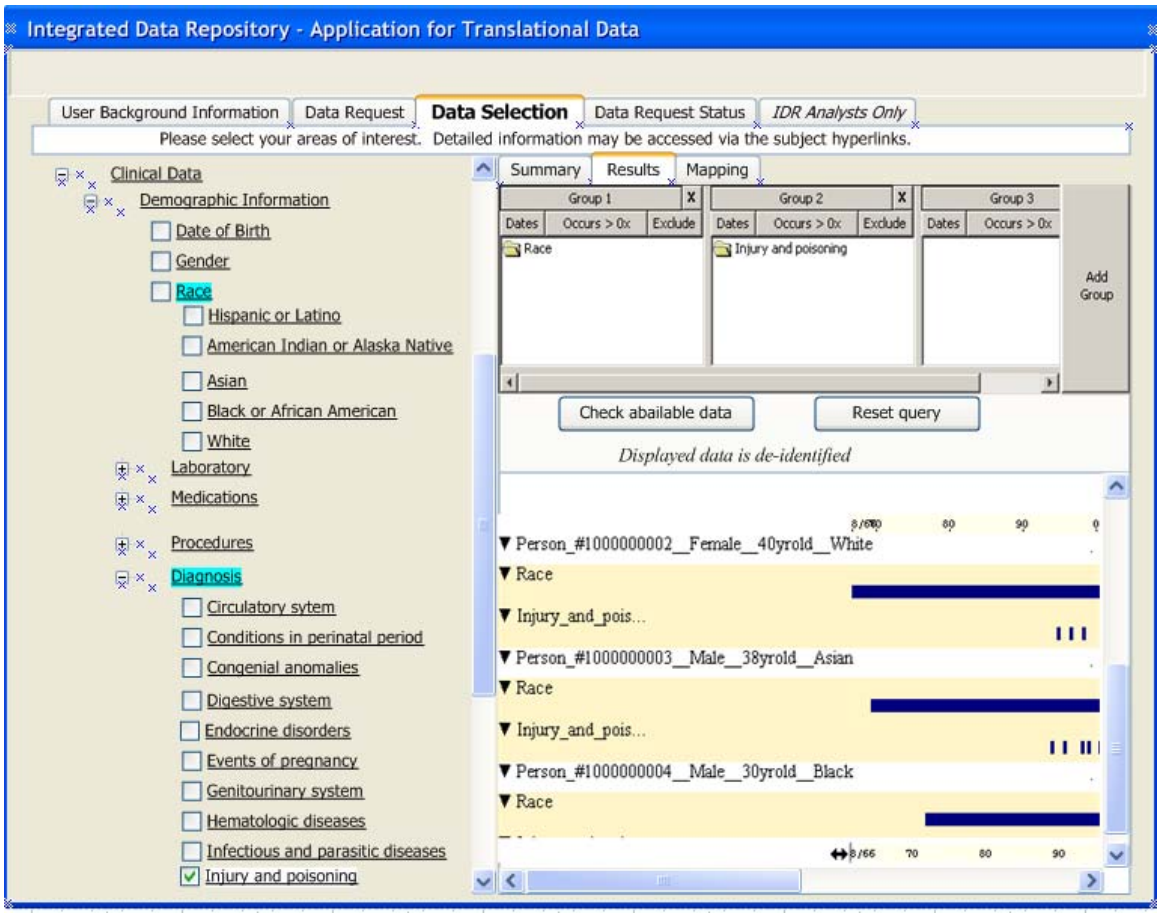
5. An approval status interface that will tell the researcher the status of the data request generated by this system.

This system will also allow a requester to browse through existing requests. That way, if another researcher has already made a similar request for data access, that work can be reused for the new request as well.

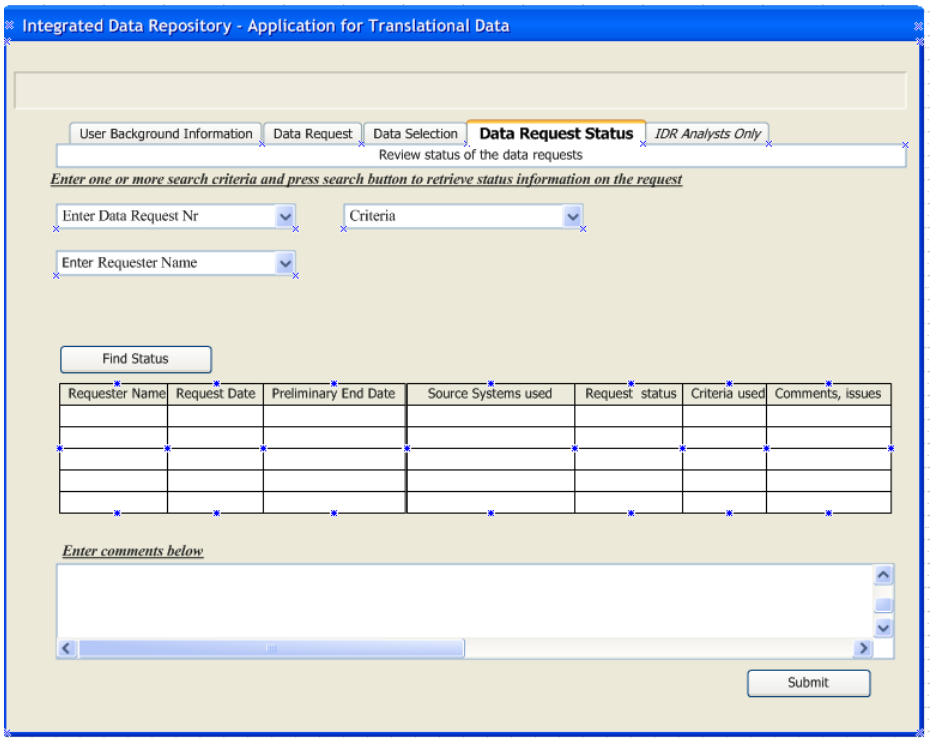


The researcher may also look at the existing types of data that are available by clicking on the Data Selection tab. This screen will enable the requester to check for available data without any assistance. If the requester comes to the conclusion that there is not enough data to support the study, the study and request can be appropriately modified or reconsidered.

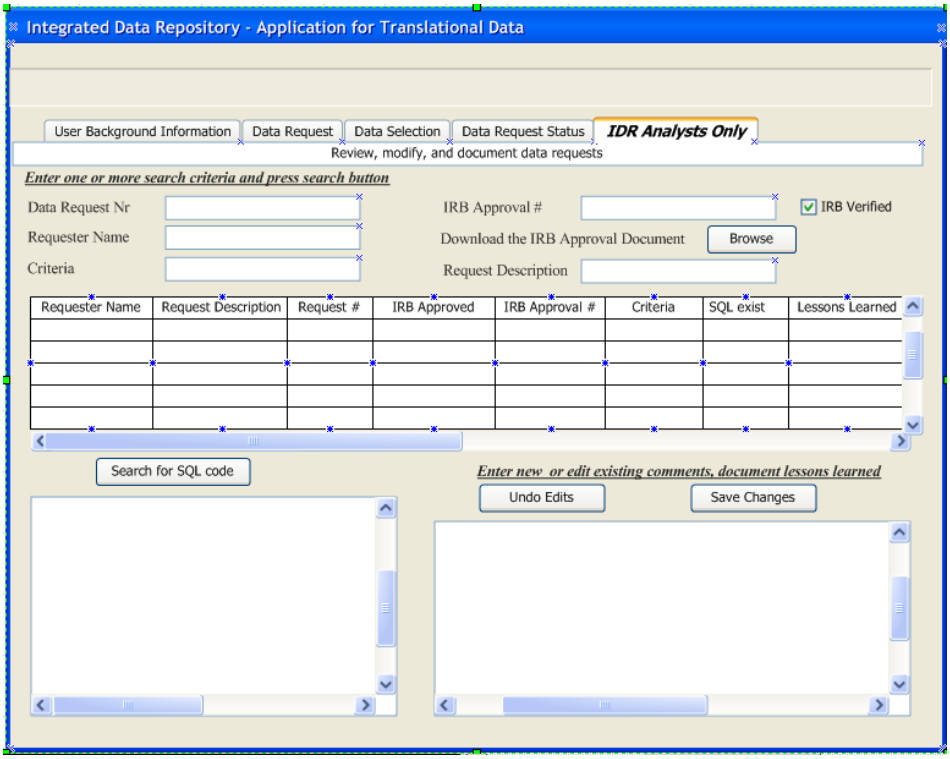
It is anticipated that this self-service system may not be able to provide all of the information required by a researcher; however, if a large portion of requesters' questions can be answered via this self-service interface, it will greatly contribute to the scalability of an IDR.



By clicking on Data Request Status, a requester can check on the progress of the request. He or she can also enter comments. Researchers may search for a request by entering either the request number, request name, or criteria.



The last tab is only accessible to Data Repository Analysts (BA or Data Analyst). Analysts may enter comments, and lessons learned as well as edit or look up existing SQL code.



This interface will be filled out by the Business Analyst. The Business Analyst will enter the IRB approval number and attach a copy of the protocol.

2.f. Inference Based Ontology Mapping

Once a researcher has determined what data elements they require, a request for access to that data must be approved by the IRB. Once access has been approved, the researcher will be given access to a view of the IDR data requested.

The researcher will have access to both the direct source data, and, when necessary, to translated (ontology mapped) data housed in the IDR. That data will have been translated by a rules-based system and be housed within “harvest tables” within the data warehouse.

2.g. Harvest Tables

The proposed tool set will *not* attempt to translate user requests for data into the researcher’s ontology of choice in real time. Instead, that data will be translated via background tasks using a rules-based system. The translated data will then be housed within harvest tables and access to those tables will be granted to the researcher’s database view. This will allow system staff to access the ontology-mapped data using standard and commercially available SQL reporting environments. There will be no requirement to construct a customized user interface for reporting and data mining that has specific knowledge of rules-based system interaction.

Additionally, ontology mapping rules developed for any particular researcher can be reused to service subsequent researcher requests for data that need to be similarly translated, via view access to the harvest tables which contain translated data.

2.h. Use of PROMPT

This proposed ontology mapping service will not attempt to fully and automatically map ontologies. Although there are initiatives that intend to offer automated mapping of medical terminology (for example, to map SNOMED CT to ICD) and the proposed system may eventually leverage that work, the present project contemplates that generating ontology maps will be a mostly manual process by a business analyst. Once a mapping has been defined it will be possible to automatically re-execute that mapping in an automated fashion as new data are imported into the repository from that same source.

This system will make the following two assumptions:

1. A particular data element imported from a specific data source will always be encoded with the same ontology.

2. The mapping of that source ontology to the ontology of use for a researcher will be fixed and, once defined, unlikely to change.

An IDR needs only to map ontologies between specific data elements contained within specific versions of fixed ontologies. In our application only a subset of the ontologies need to be mapped and that mapping is fixed based on the source of the data and the version of the ontology, and does not change once configured.

If the encoding of data that is imported into the IDR must be altered, the system owners of the IDR must be notified in advance. However, if a data definition were to change without prior notification to the IDR system owners, the system will detect that change in encoding and generate a system alert error message.

Although application of this technology to the Semantic Web is clearly out of scope for this project, this same technology could later be used to provide enhanced Semantic Web technology for the IDR by offering an alternative approach to the ad-hoc mapping of ontologies.

The Ontology Mapping Service will provide the following specific features:

- Generation of ontology-mapped data.

- Mapped data will be stored in harvest tables within the IDR that are compatible with standard SQL reporting and data mining tools.

- The ability to store mapping system rules within the IDR and subsequently refer to those rules to perform ontology mappings as a background task. (The selection of the mapping to perform will be made on the Mappings Tab shown in Figure 3 above.)

- The ability to transmit email alerts which indicate that a specific rule has been triggered. This can be useful when a researcher needs to know immediately when new data has been imported and made available within the IDR.

The proposed tools will provide ontology mapping via the PROMPT system, which is part of the Protégé knowledge management system [2]. Three components of PROMPT will form the basis of this implementation:

1) The Protégé Prompt Knowledge Mapping Tool
This software client has been created for the specific purpose of creating mapping rules of the sort required for this project.

2) The Mapping Interpreter

This is a command line tool capable of mapping data from source to target ontologies using a list of mapping relations. These mapping relations are instances of the Ontology of Mapping Relations. Although this interpreter does already exist within the Prompt environment, it will require some additional engineering for use within the Ontology Mapper.

3) The Ontology of Mapping Relations [1]
This preliminary work is an ontology designed for the specific purpose of mapping data encoded in other ontologies. We intend to extend this ontology to include the formalism required for development of an Ontology Mapper.

To successfully incorporate these three components the Ontology Mapping Service will provide the following features:

The ability to import ontology mapping relations created using the knowledge mapping tool, and subsequently select those rules for execution via the mapping tab shown in Figure 3 above. The mapping interpreter will run as a background service and perform the selected mapping functions.

Specifically, the system will:

1) Detect when new data has been imported into the IDR.

2) Examine meta-data associated with the imported data to see if particular data elements have been updated.

3) Determine if there are ontology mapping relations which need to be executed on any of the newly updated data.

4) Read in the source data and for each element, call the mapping interpreter to convert the data into its target format.

5) Write the new data into the associated harvest table so that subsequent reporting can be performed on the results.

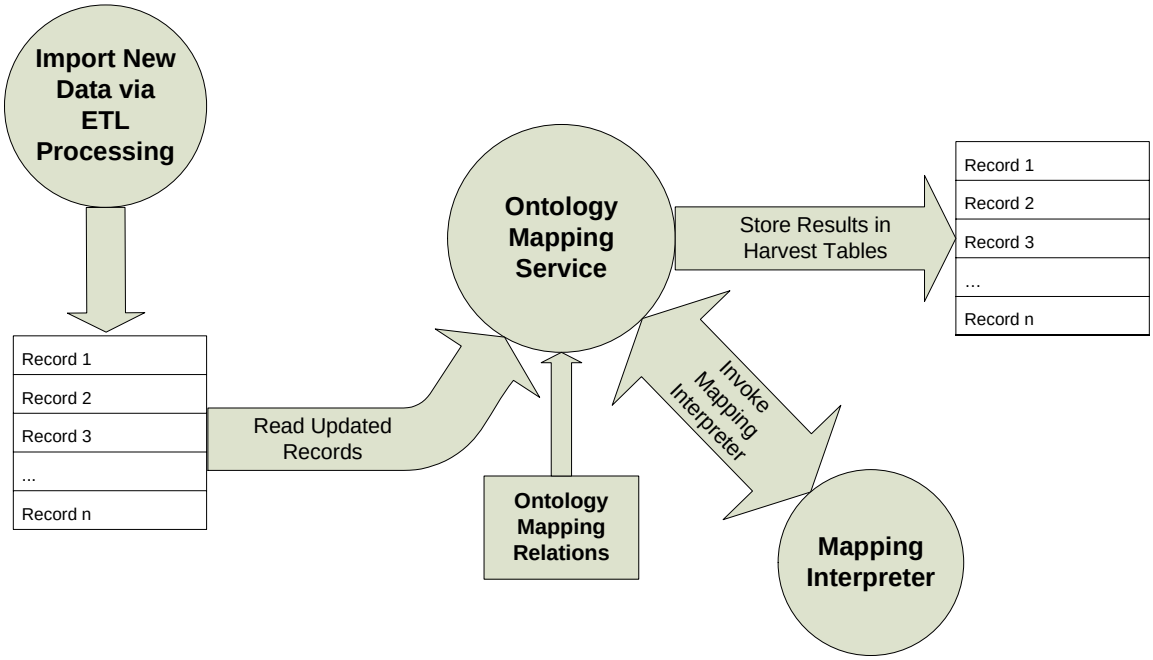


Figure 7. *The major architectural elements of the Ontology Mapping Service*.

The network protocol used to access this service will support the HL7 CTS II mapping standard.

In order to accomplish these goals the Ontology Mapping Service will need to implement several new software functions:

- An import feature capable of reading in ontology mapping relations that have been created using the PROMPT Knowledge Mapping Tool [3].

- A background service capable of calling the PROMPT mapping interpreter to map specific data items. This interpreter will read source data and apply the associated mapping relation to generate the result set.

- A means of describing the source data (Figure 2, above), and the requested ontology mapping (Figure 3, above), and the automatic generation of a harvest table within which to store the results of the mapping.

- The ability to access multiple instances of the Ontology Mapping Service including commercial mapping services that will provide mapping services via HL7 CTS II over the internet.

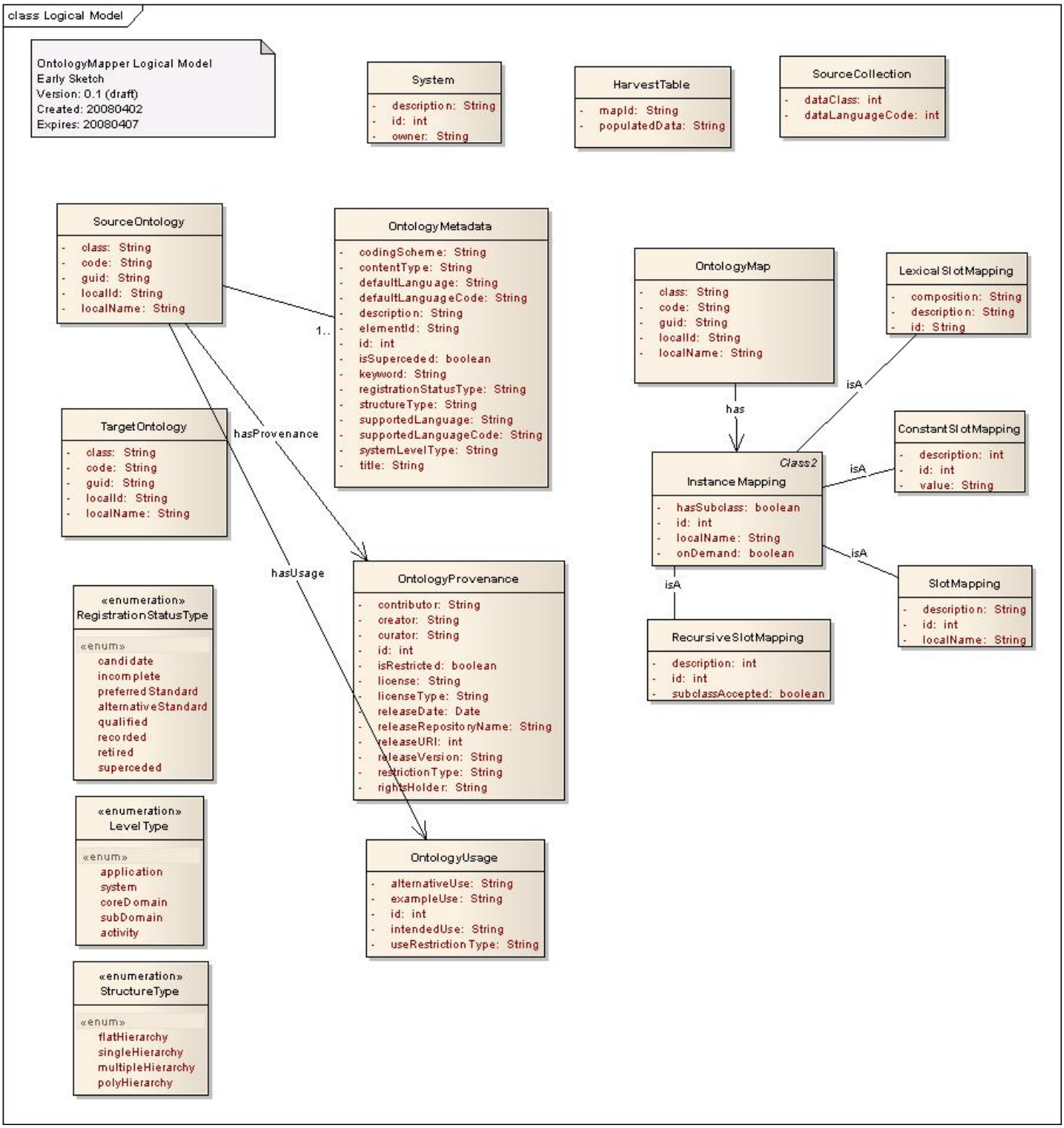
2.i. Mapping Relations

For the mapping interpreter to work, it must have a formal set of possible mapping relations that it applies to convert data. We have chosen to use an ontology of mapping relations, as defined by Crubezy et al. [1], as a starting point for this formalism. We realize that (a) this work is limited to mapping relations between objects in a frame-based knowledge representation formalism, and (b) this ontology has not been extensively tested with data that are primarily stored within a relational database, and that (c) sometimes data are not encoded in a formal ontology but instead exist as a standard terminology. Nonetheless, we are confident we can modify and rebuild this formalism to suit our needs.

2.j. HL7 CTS II

During the initial design of this proposal our group made contact with the HL7 CTS II Technical committee in charge of data interchange specifications. The CTS II committee expressed a strong interest in incorporating the exchange of ontology mapping services into their HL7 CTS II specification. We have therefore supplied HL7 with our planned interface specification for the Ontology Mapper for inclusion into the new CTS II specifications due for publication this year.

This system will use a *subset* of that new HL7 CTS II specification for communications with the Mapping Interpreter. This new interface will allow data mapping using the local mapping interpreter or via subscription to commercial mapping services over the internet. To accomplish this the Mapping Interpreter and the commercial services specification included in CTS II will use the same network protocol. The data model used for communications with the mapping interpreter is described in the following diagram.



The logical data model includes work that features work recently developed by the caBIG community for terminology metadata as well as modeling derived from work by Crubézy et al. At the center of these structures are Metadata, Provenance, and System tables that address high-level administrative and data ownership information requirements. This includes: 1) metadata for provenance with institutional metadata; 2) locally and globally unique and human-readable object identifiers for all objects and actors, including those who are entities responsible for the mapping (e.g. creator); 3) individuals contributing or performing the activity (e.g. contributors); and 4) those with primary responsibility such as oversight or review (e.g. curators). Each mapping will intrinsically have a source and a target instance and for every instance will require a robust set of

attributes to uniquely identify the map both locally and globally. These data elements will also provide information regarding map derivation and details about the nature of the transformation activity.

The maps, relationships, and data transform structures are represented by the Ontology Map and mapping tables. Relationships or associations (including collections) will have their own set of metadata such as unambiguous descriptions, directionality, cardinality, etc. Although this diagram shows data elements with enumerated value domains, those listed are suggestions for development in this early model. Maps will have associated identifiers not only about themselves, but also their relationship to a Harvest table. MapRules are textual data that contain an XML encoded mapping rule.

2.k. CTS II Interaction

The following network requests may be made via HL7 CTS II in order to populate the user interface of the Discovery Interface. The network requests may either be made to the local Ontology Mapping Service or to a commercial mapping service that follows this network protocol.

1. Identify Available Mappings

Returns a listing of terminology mappings on the specified instance of the terminology service. The Discovery Interface may request a list of available mappings.

2. Resolve Mapping Reference Information

The Discovery Interface may request all metadata associated with a specific mapping.

3. Resolve Available Code Systems

Provides a listing of all of the different coding systems available from the terminology service.

4. Map Individual Element and Resolve Element Reference Information

- Request all available System IDs
  The Discovery Interface may request all available System IDs and associated metadata.

- Request Source Element IDs for Specific System ID
  The Discovery Interface may request all source Element ID information.

- Request Target Element IDs for Specific System ID
  The Discovery Interface may request all target Element ID information.

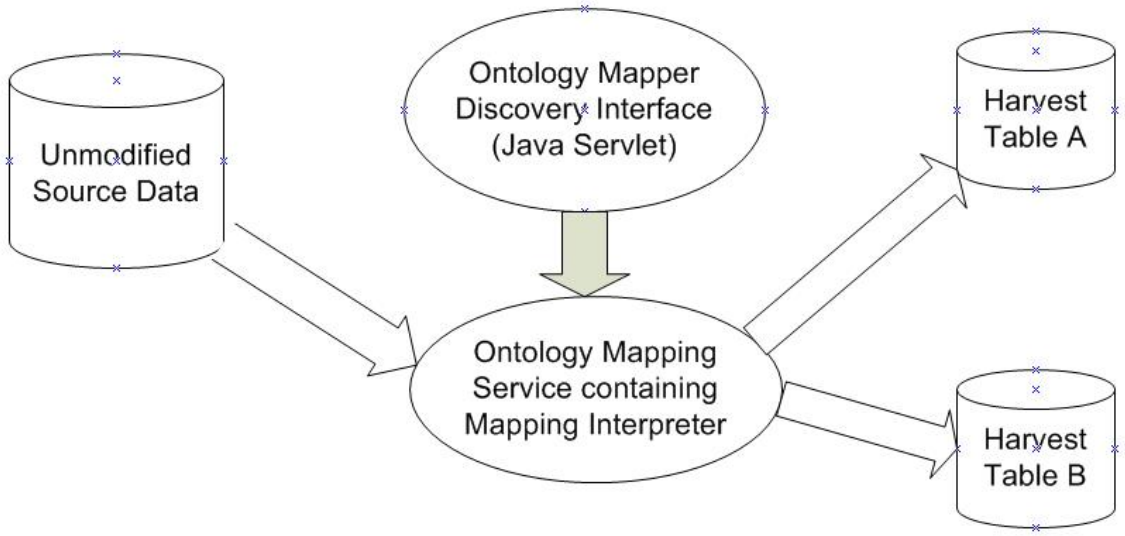
- Request Mapping Rule Content
  The Discovery Interface may request the XML rule that was encoded to describe how a specific mapping is to be performed.

Once this information has been collected the local Mapping Interpreter will be configured to run as a background process to perform the mapping of data as described in the Rule Content.

The CTS II Interaction model involves the transmission of the mapping rule and its associated metadata. The actual data mapped is not transmitted via this protocol as data is always mapped using local computational resources.

2.l. Runtime Components

This system would consist of only two runtime components, an Ontology Mapper Discovery Interface that accepts and tracks user requests and an Ontology Mapping Service and its associated Mapping Interpreter. This service would run as a background task and process data according to a preconfigured schedule.



2.m. Rule Types

There will be several different types of rules encoded into this system:

- Ontology Mapping: translates data from one standard encoding to another
- Fact Extraction: creation of new information based on the inputs to the system.
- Prototype Selection: The ability to assign a mode to the dataset which can be used in queries and subsequent rules. For example, even though there may not be any single data element within the IDR to indicate that a person is disabled, a Prototype Selection rule could be used to generate a new data element to indicate that the "Person Is Disabled", thereby greatly decreasing the complexity of future queries that rely on this same state.
- Quality Control: The detection of conflicts for follow-up via corrective action procedures. For example, a lab test code fails to map to a formal ontology.
- Aggregation Generation: The ability to roll up a dataset based on a particular dimension. For example, "provide a summary of all cases of congestive heart failure, based on its ontology mapped term, for a given time period."

2.n. Security Plan

Integrated Data Repositories pose special challenges regarding information security and compliance. This proposal will implement the highest standards of security, while balancing privacy and compliance with the need for convenient data access by authorized personnel.

Specifically we shall:

- Encrypt at rest all Protected Health Information as defined by the regulations of the Health Insurance Portability and Accountability Act (HIPAA), and encrypt any other data requiring enhanced security
- Implement FISMA security guidelines for access to any data obtained via an agreement with a Veterans Administration health facility
- Restrict researcher access to data in accordance with the regulations of the Office of Human Research Protection and associated IRB approval processes

- Maintain detailed audit reports of end user activity
- Implement a monthly password rotation policy for all administrative level access
- Encrypt all data transports via SSL using HTTPS and SFTP
- Require all users to sign a security practices agreement

Conclusion

Our proposed design is intended to compensate for the high initial cost and resource allocation required with up-front data governance. Through the use of a rules-based system the translation of data into the domain of a specific researcher can be accomplished more quickly and efficiently than with a traditional data warehouse design. The proposed system will dramatically lower the barrier to IDR development at individual CTSAs to support translational science, and promote inter-CTSA data sharing and research collaboration.

References

1. Crubezy M., Pincus Z., Musen M. Mediating Knowledge between Application Components. Stanford University

2. J. H. Gennari, M. A. Musen, R. W. Fergerson, W. E. Grosso, M. Crubezy, H. Eriksson, N. F. Noy, and S. W. Tu. The Evolution of Protégé: An Environment for Knowledge-Based Systems Development. International Journal of Human Computer Studies, 58(1):89-123, 2003

3. Noy NF, Musen MA. The PROMPT suite: Interactive tools for ontology merging and mapping. *Int J of Human-Computer Studies* 2003;59(6):983-1024.\

4. Novak J, Canas A. The Theory Underlying Concept Maps and How to Construct Them. Technical Report IHMC CmapTools 2006-01

5. Gardin, J., Schumacher, D., et al. (2000). Valvular abnormalities and cardiovascular status following exposure to dexfenfluramine or phentermine/fenfluramine. *Journal of the American Medical Associatio*n, 283: 1703-1709

6. Advani A, Tu S., OConnor M., Coleman R., Goldstein, M., Musen M. Integrating a Modern Knowledge-Based System Architecture with a Legacy VA Database: The ATHENA and EON Projects at Stanford

7. Judith J. Warren, Jeanne Collins, Catherine Sorrentino, James R. Cambell. Just-in-Time Coding of the Problem List in a Clinical Environment: Univ. of Netbraska Medical Center, Omaha, NE

3. PROJECT MANAGEMENT AND STAFFING

3.a. Approach to Project Management

This proposal represents a collaborative effort between the CTSA's located at the University of California, San Francisco (UCSF), University of California, Davis (UC Davis), University of Pennsylvania, University of Washington and the University of Rochester. UCSF is the largest CTSA site with considerable institutional backing and resources, and will provide the leadership and overall project management required to complete this project. This consortium will have six teams that will concentrate on the specific implementation goals of the project. Members of these groups overlap based on prior experience to further enhance interdisciplinary interaction between team members.

1. Project Oversight: This team will focus on maintaining a clear project vision, track overall project status and will review the project management reports gathered from the other teams. Another responsibility of this group will focus on end-user Use Case encoding to maintain clear guidelines on the deliverables.

2. Overall System Design: This team will focus on establishing the overall system design and construction necessary to integrate the other components of this environment. This team will also maintain a regression test suite that will be composed of unit tests provided by the other teams.

3. Rules-Based System: This team will be responsible for the design and building of the rules-based system at the heart of the Ontology Mapper. This will include both the extensions required to the Ontology of Mapping Relations, the enhancements to the Protégé Prompt knowledge management system, and for the construction of the mapping interpreter servlet.

4. Data Warehouse Database Design: This team will be responsible for the construction of the database schema and metadata storage required for the operation of the Ontology Mapper. This group will also be responsible for the design and construction of the Harvest Table component and will build the components required for identifying when new data has been imported into the IDR and will require translation as a background service.

5. HL7 Interface: This team will be responsible for the construction of the HL7 CTS II interface, which will incorporate our contribution to the CTS II specification for the transmission of mapping rules. (A subset of CTS II described above which has already been accepted by the HL7 CTS II committee.) This team will be responsible for those components that can subscribe to CTS II based mapping services and for the construction of that service within the mapping interpreter.

6. Data Request Lifecycle: This team will be responsible for the Data Discovery, Data Request, and Lifecycle Tracking portions of the Ontology Mapper User Interface. They will work closely with the Data Warehouse team to help describe how that metadata will be described within the database.

3.b. Collaborative Software Development

Collaborative software development projects can offer special challenges due to the nature of distributed project management over a wide geographic area. To overcome those obstacles it is necessary to foster deep social interactions and trust between team members. Team members will hold frequent team meetings to identify the immediate goals for each individual on the team. These repeated meetings will test the credibility of each team member and provide management with the feedback necessary to immediately address roadblocks. These team development meetings will be held over UCSF's advanced video conference system to as closely approximate face-to-face interaction as is possible in a distributed environment. The project management methodology followed will be Agile Software Development as described by the following five levels.

Level 1 - Product Visioning
At all times a clear statement of the goals of this project will be maintained.

Level 2 - Product Roadmap
Although we have created and will maintain a long-term project schedule, the software development roadmap for the Ontology Mapper will instead be based on high frequency and small deliverables, which are measured in days or weeks. These smaller deliverables will contribute towards the project schedule in a cumulative effect.

Level 3 - Release Planning
The initial product release and major software components required for the delivery of the Ontology Mapper have been outlined in this proposal. Those components have been used as the basis for developing

specialized software development groups that will focus on specific aspects of this environment based on past experience and resource availability.

Level 4 - Iteration Planning
There will be several iterations of software development as each small scale deliverable is defined, designed, implemented, integrated, and tested. These iterations will be aligned with the development of specific deliverables. Prior to beginning work on each of these iterations, an iteration planning session will be held in order to verify that the tasks of the group are matched by that group's capacity to deliver the required software. The team involved with each iteration will be responsible for creating their own unit tests, which will be integrated into an overall regression test suite by the System Design Team. Specifically, each Iteration Plan will contain the following:
1. List of the minimum number of features needed in priority order for the iteration.
2. Component development efforts are broken into 15-day iterations (called Sprints).
3. Each team will contain all the expertise needed to complete the iteration.

Level 5 - Daily Plan
Each day a very short planning session will be held for one of the development groups. This session will be used to review the work ahead and any issues that need to be addressed. Regular meetings of this sort are essential to establish trust within the group and to enhance team commitment within a distributed software development environment.

3.c. Staffing Plan

All available resources have been sorted by the required skills below. These groupings form the basis of our release and iteration planning as described above and will be augmented to include required Java programming expertise. In order to meet the deliverables of the project these positions will be augmented with contracts to hire as specified in the budget details for each of the contracting sites.

Administrative Oversight

Russ Cucina – UCSF Principal Investigator
Rob Wynden – UCSF Project Director
Jonathan Showstack – UCSF Co-Investigator
Michael Kamerick – UCSF Key Personnel
Stuart Turner – UCD Principal Investigator
Mike Hogarth – UCD Co-Investigator
Marco Casale – University of Rochester Co-Investigator
David Krusch – University of Rochester Principal Investigator
Maggie Massary – University of Pennsylvania Project Director
Mark Weiner – University of Pennsylvania Principal Investigator
Mark Musen – Ontology Mapper Project Key Personnel
Russ Ham – Ontology Mapper Project Key Personnel

Project Management

Rob Wynden – UCSF Project Director
Gail Harden – UCSF Project Manager
Davera Gabriel – UCD Project Director

Overall System Design

Rob Wynden – UCSF Project Director
Stuart Turner – UCD Principal Investigator

Rules Based System

Russ Cucina – UCSF Principal Investigator
John Gennari – University of Washington Contractor Resource
Rob Wynden – UCSF Project Director
Stuart Turner – UCD Principal Investigator
Davara Gabriel – UCD Project Director (Terminologist for Standards Harmonization)

Data Warehouse Database Construction

Rob Wynden – UCSF Project Director
Marco Casale – University of Rochester
William DiGrazio – University of Rochester
Maggie Massary – University of Pennsylvania Project Director

Clinical Expertise

Russ Cucina – UCSF Principal Investigator
David Krusch – University of Rochester Principal Investigator
Mark Weiner – University of Pennsylvania Principal Investigator
Mike Hogarth – UCD Co-Investigator

Terminologists

Davera Gabriel – UCD Project Director
Stuart Turner – UCD Principal Investigator

HL7 Expertise

Davera Gabriel UCD Project Director (Standards Harmonization)
Marco Casale – University of Rochester Co-Investigator
Russ Hamm – HL7 CTS II Advisor Resource
Rob Wynden – UCSF Project Director
Stuart Turner – UCD Principal Investigator

Data Discovery and Data Request Lifecycle Tracking

Maggie Massary – University of Pennsylvania Project Director
Davera Gabriel – UCD Project Director
Rob Wynden – UCSF Project Director

4. DOMAIN KNOWLEDGE AND EXPERIENCE

Summary

Each of our respective institutions brings to this project a highly valuable set of technical skills. Together we are uniquely suited for the task of creating the Ontology Mapper.

UCSF
- Administrative Oversight
- Project Management
- Ontology Mapper System Design
- Data Warehouse Construction
- Clinical Expertise
- IDR Security

UC Davis
- Terminologist
- Standards Harmonization
- System Design
- HL7 Expertise
- Clinical Expertise

U Pennsylvania
- Project Lifecycle Tracking
- Data Warehouse Construction
- Clinical Expertise

U Rochester
- Data Warehouse Construction
- HL7 Expertise
- Clinical Expertise

5. DATA AND SOFTWARE SHARING PLANS

The results of the work completed under this grant application will be distributed for use throughout the CTSA. To assist in the dissemination of this technology to CTSA member sites the following initiatives will be pursued.

1. Open Source Lesser GPL (LGPL) software will be shared CTSA-wide.

The software developed under LGPL license (http://www.opensource.org/licenses/lgpl-3.0.html) and the full source code for the work product will be made available on SourceForge.net.

2. Mapping Rules Registry Service available on the internet.

It is our belief that the rules sets required for mapping terminologies are themselves a valuable and shareable framework. The Ontology Mapper project will incorporate both a standardized method of encoding those rules and a standardized method of subscribing to those rules over the internet. To facilitate sharing of mapping rules specifically within the CTSAs, we will create a Rules Registry website. The Rules Registry will allow CTSA member sites to post Ontology Mapper rules sets, fostering collaboration by CTSA sites on the development and maintenance of these rules.

3. Ontology Mapper User Group.

An Ontology Mapper user group will be created to assist in the education of CTSA sites interested in the implementation of this environment. We will also use the user group as a means of collecting user feedback required for ongoing system enhancements.

4. Facilitate the use of this technology within initiatives which focus on the secure network transmission of data between CTSA sites, such as the CTSA Human Studies Metadata Repository and the I2B2 Federated Query mechanism.

There are other technology initiatives within the CTSA which are focused specifically on the transmission of data between CTSA sites. The data delivered from a group of IDR implementations would not likely be encoded in a common format and therefore the Ontology Mapper project may assist these efforts. This Ontology Mapper could be used to encode data on either the source or the target end of a collaborative research network so that the translated and aggregated information from all sites can then be utilized via a common terminology.

5. Facilitate sharing of this technology with other open source informatics platforms, such as the I2B2 Platform

There exist open source translational informatics platforms which would benefit from the availability of this technology. The Ontology Mapper group will actively seek to engage those platforms in the creation of modules which would allow the direct linkage of the Ontology Mapper to those environments. The Ontology Mapper group is already aware of just such an effort under way within the I2B2.org community.

Data Sharing Use Cases

The following use cases illustrate how the proposed environment may be used for sharing data among institutions in the context of an Integrated Data Repository.

Primary Actor	***Use Cases***
Clinical Investigator (CI)	CI Cross Site Collaboration
Clinical Investigator (CI)	Search for Aggregated clinical study data, Create Custom Mapping
Integrated Data Repository (IDR)	Search for Aggregated clinical study data, Create Custom Mapping
Business Analyst (BA)	Create Custom Mapping
Protégé Prompt	Create Custom Mapping

Use Case #1 – Clinical Investigator Cross-Site Collaboration

Actors:	Clinical Investigator (CI), Integrated Data Repository (IDR)
Description:	A clinical investigator is attempting to collaborate with another clinical investigator who works within a partner institution.
Trigger:	A Clinical Investigator is performing a collaborative study on the drug Avandia.
Preconditions:	The integrated data repository available to the CI has the required data but that data must be translated into the correct terminology and aggregated as appropriate.
Postconditions:	The Clinical Investigator obtains an aggregated de-identified data, and following IRB approval, transmits that information to the partner institution via a secure network protocol.
Normal Flow:	Use Case 1 is Triggered. The aggregated data is then transmitted via a secure network protocol to the partner institution.

Use Case #2 – Search for Aggregated Clinical Study Data

Actors:	Clinical Investigator (CI), Integrated Data Repository (IDR)
Description:	A clinical investigator would like to search for all available aggregated clinical study data regarding the use of the drug Avandia and its impact on cases of myocardial infarction (MI).
Trigger:	A Clinical Investigator is performing a study on the drug Avandia and a comparative study of related weight loss drugs for diabetes to determine if there is a common impact on cases of MI
Preconditions:	The integrated data repository has both cases of MI and medication orders of Avandia or similar medications but does not have a direct relationship between them.
Postconditions:	The Clinical Investigator obtains an aggregated de-identified data set regarding clinical trials or any patient encounters from the EHR on the use of Avandia and similar medications with any association related to MI. The data set is composed of variables from the IDR which are derived from electronic health record data, clinical trials data and genomics data.
Normal Flow:	1. The Clinical Investigator logs into the CTSA Ontology Mapper Application 2. The Clinical Investigator requests a list of pharmacy ontology maps. 3. The IDR's Ontology Mapper sends a request to an HL7 CTS II terminology service requesting available pharmacy ontology maps. 4. The IDR Ontology Mapper receives a list of available pharmacy maps including : RxNorm, SNOMED-CT, Medi-Span GPI (Generic Product Identifier) 5. The Clinical Investigator selects Avandia from a custom built ontology for the clinical investigator's healthcare system and requests a map from the custom built medication ontology to RxNorm, SNOMED-CT and GPI 6. The Ontology Mapper mines all data that uses the selected drug term (custom ontology) along with any of the ontology cross mapped terms. 7. The Ontology Mapper Aggregate Generator Rule is triggered to summarize the result set into an aggregated list of total clinical trials performed on Avandia with associated EHR and genomics variables associated with the study cohort. 8. The Clinical Investigator extracts the data set.
Alternative Flows:	The Clinical Investigator selects an alternative drug that is similar to Avandia for comparative analysis.
Exceptions:	The Ontology Mapper is notified that there are no available maps. The clinical investigator can only use the internal custom ontology maps for the drug. Use Case 2 is triggered.
Includes:	a. Map Individual Element and Resolve Element Reference Information. The Ontology Mapper requests a HL7 CTS II Ontology Mapping Service to map the custom medication ontology to RxNorm. b. Map Individual Element and Resolve Element Reference Information. The Ontology Mapper requests a HL7 CTS II Ontology Mapping Service to map the custom medication ontology to SNOMED-CT c. Map Individual Element and Resolve Element Reference Information. The Ontology Mapper requests a HL7 CTS II Ontology Mapping Service to map the custom medication ontology to Medi-Span GPI
Business Rules:	Obtain current clinical research information for a projected research project

Use Case #3 – Create Custom Mapping

Actors:	Clinical Investigator (CI), Business Analyst (BA), Integrated Data Repository (IDR)
Description:	A clinical investigator would like to have a custom ontology mapping creating in the IDR.
Trigger:	A CI has a need for a custom ontology mapping
Preconditions:	The IDR does not contain a custom ontology mapping for the specified terms requested by the CI
Postconditions:	The CI is able to receive terms that are crossmapped in the IDR
Normal Flow:	1. The CI logs into the IDR User Interface (UI) 2. The CI requests that the new custom mapping be created (using the 'Other' text box in the UI) 3. The BA would manually inspect the IDR to determine which table has the data in it (although it has not been encoded correctly) 4. The BA would use Data Discovery User Interface to determine any special notes or contextual information about the data source 5. The BA logs into the Protégé Prompt application 6. The BA loads the extended version of the Ontology of Mapping Relations 7. The BA would create the new map in Protégé Prompt 8. Protégé Prompt generates a new XML file for the new mapping 9. The BA logs into the IDR UI and uploads the XML file into the IDR 10. The BA logs verifies the new mapping by selecting the new mapping within the IDR UI.
Business Rules:	Obtain current clinical research information for a projected research project

6. INVENTORY OF BIOGRAPHICAL SKETCHES

Prime Respondent:
University of California, San Francisco (CTSA funds recipient)

Principal Investigator:
Russell J. Cucina, MD, MS, Assistant Health Sciences Professor of Medicine and Center for Clinical and Translational Informatics, University of California, San Francisco and UCSF Clinical and Translational Science Institute (CTSI); Associate Medical Director of Information Technology, UCSF Medical Center, San Francisco, CA

Primary Site Co-Investigators:
Jonathan Showstack, Ph.D., Assistant Vice Chancellor, Academic and Administrative Information Systems; Adjunct Professor of Medicine and Health Policy, University of California, San Francisco and UCSF Clinical and Translational Science Institute (CTSI)

Project Director:
Rob Wynden, System Architect and Team Lead, CTSA Integrated Data Repository Project, Clinical and Translational Science Institute (CTSI), University of California, San Francisco, CA (415)476-3728

Key Technologist:
Informatics Expert Advisor: Mark A. Musen, M.D., Ph.D, Professor of Medicine (Biomedical Informatics); Division Head (BMIR); Co-Director, Biomedical Informatics Training Program

HL7 Expert Advisor: Russ Hamm, Project Lead for HL7 CTS II, Co-chair of the HL7 Vocabulary Technical Committee, Informatics Specialist at Apalon Inc.

Data Governance Advisor: Michael Kamerick Director, Academic Research Systems, Office of Academic & Administrative Information Systems, UCSF

Partnering Organizations:
1. Principal Investigator: David A. Krusch MD, University of Rochester Medical Center
co-Investigator: William DiGrazio MS, MBA, University of Rochester Medical Center
co-Investigator and Project Director: Marco Casale MS, University of Rochester Medical Center

2. Principal Investigator: Stuart Turner DVM, University of California, Davis, CA
Co-Investigator: Michael Hogarth, MD, Associate Professor in both the Department of Internal Medicine and the Department Pathology in the UC Davis School of Medicine
Project Director: Davera Gabriel RN, University of California, Davis, CA

3. Principal Investigator: Mark Weiner, MD, Dir of Info Sys Integration for Research,
University of Pennsylvania School of Medicine
Project Director: Maggie Massary, Clinical Informatics Manager, University of Pennsylvania Health System

4. Principal Investigator: John Gennari, PhD, Associate Professor, Department of Medical Education and Biomedical Informatics, University of Washington