Inter-institutional data sharing: the CICTR project

Dec 8, 2009

Nick Anderson, Ph.D.
Assistant Professor, Biomedical Health Informatics
University of Washington
Institute of Translational Health Sciences

ITHS
Institute *of* Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics

Overview

- Scope of project and current status
- Technical environments
- Challenges and next steps

Scope of Project

- Two year demonstration project (11/08-11/10)
- Three CTSA partners with academic medical centers
  - University of Washington ITHS
  - UC San Francisco CTSI
  - UC Davis CTSC
  - Harvard Catalyst (collaborator)
- Private data warehousing company Recombinant Data Systems
- www.i2b2.org

ITHS
Institute of Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics

Project Goals

- Foster data driven research collaborations
- Develop and test generalizability of using anonymized data to support federated querying across geographically distributed academic institutional medical systems
- Evaluate impact of systems and processes on different classes of end-users and institutions
- Pilot governance approaches to support/protect patients, researchers and institutions

Use case: Cohort discovery

- Q: How many patients in the UWMC system might be at risk for diabetes?
- Inclusion criteria
  - Ages 18-40
  - Obesity (ICD-9 278.*)
  - Other abnormal glucose (ICD-9 790.29)
- Exclusion criteria
  - Diabetes Mellitus (ICD-9 250.*)

Use case writ larger…

- How can (or just can) this query be parsed against external clinical populations?
- How can sensitivity and specificity be increased?
- How can these results be effectively used?

Multi-Institutional Use-cases and Users

- Anonymized cohort discovery for clinical trial recruitment
  - Current: aggregate counts and institutional source
  - Future:
    - » Descriptive metadata
    - » Local HIPAA de-identified Limited Data Sets
- Intended users:
  - Clinical translational investigators/study teams
  - Informaticians
  - Terminologists
  - Public health researchers (pending)

Four parallel processes

- Technical - IT/development/implementation/testing
- Governance - Data Use Agreements/IRB institutional alignment
- Ontology - Terminologies/semantic alignment
- Evaluation - Process, outcomes and usability evaluation

Human resources needed

- Formal (e.g. – paid)
  - Informaticians
  - ETL analysts
  - Terminologists
  - Software architects/Developers
  - Usability/Evaluation researchers
  - Informatics/Information science students
  - IT staff
- Informal (e.g. – priceless)
  - Support of project at highest institutional and regulatory levels - CIO/CTO
  - (new!) clinicians/clinical researchers

Technical environments

- Compatible server architectures/different DB environments architecture environments
- “Identical” I2b2 environments
- Common ETL and anonymization processes
- Common development environment
- Common knowledge environment
- Broadly similar governance processes

I2b2

- Informatics for Integrating Biology and the Bedside (www.i2b2.org)
- NCBC funded center - grew out of RPDR (Harvard, Partners Health, Mass Gen)
- Multiple biological cores – 1 core is software
- Deployed 29+ institutions world wide
- Implemented as web services against Oracle/SQL server/Sybase IQ
- Java/LAMP v1.3 (1.4 this month)
- GPL license

I2b2 Hive Environment

I2b2 Workbench

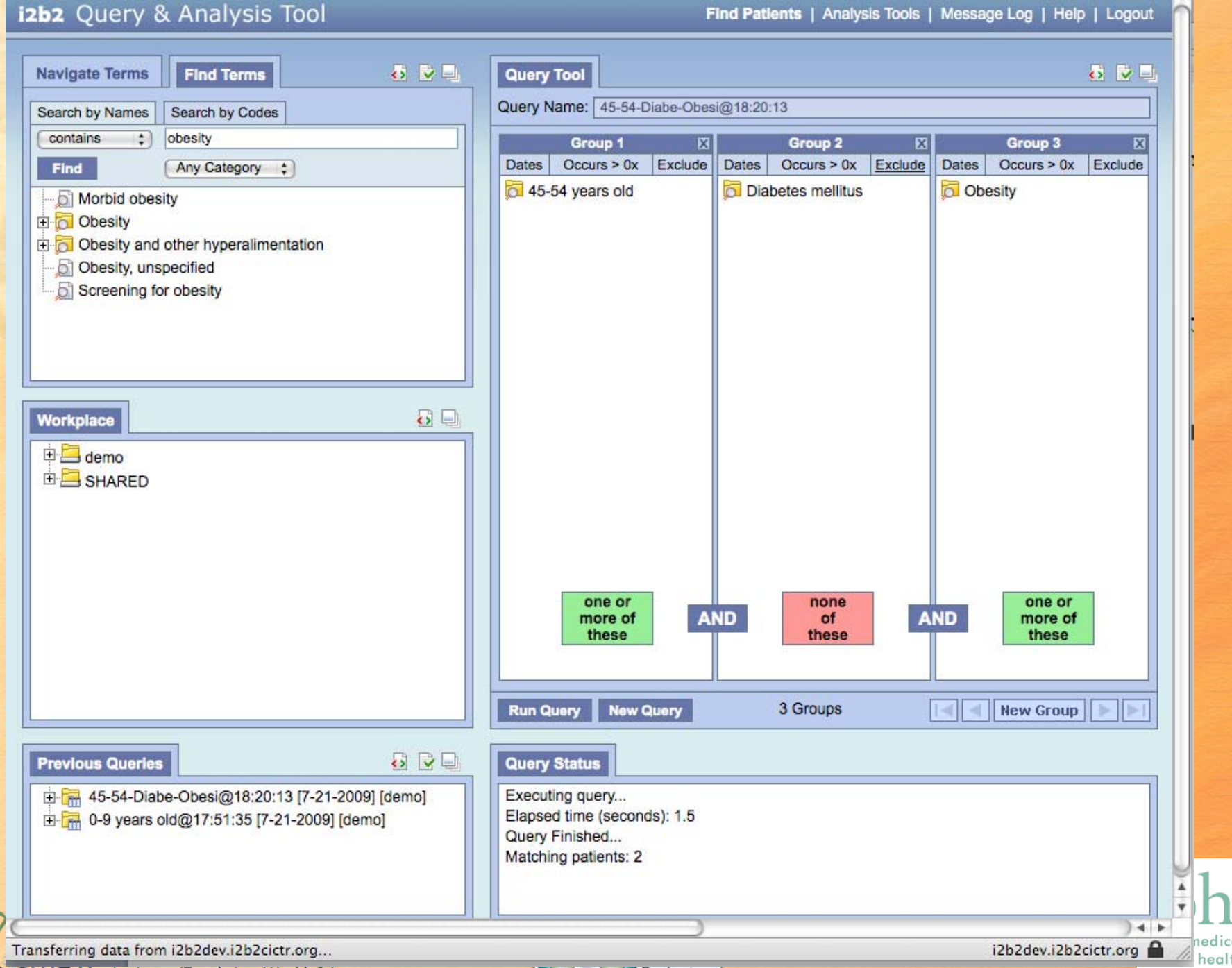
Client side (desktop or web)

(mostly) server side

ITHS Institute of Translational Health Sciences

Information Exchange Pilot Project

bhi biomedical and health informatics

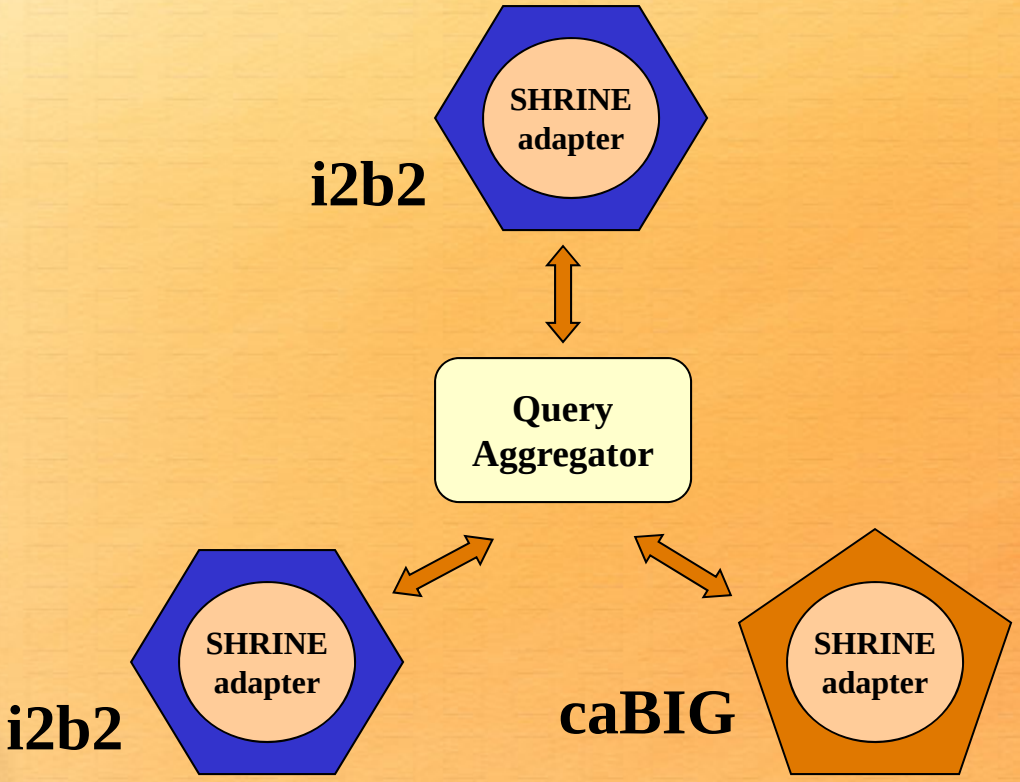


Institute of Translational Health Sciences

Project

medical health informatics

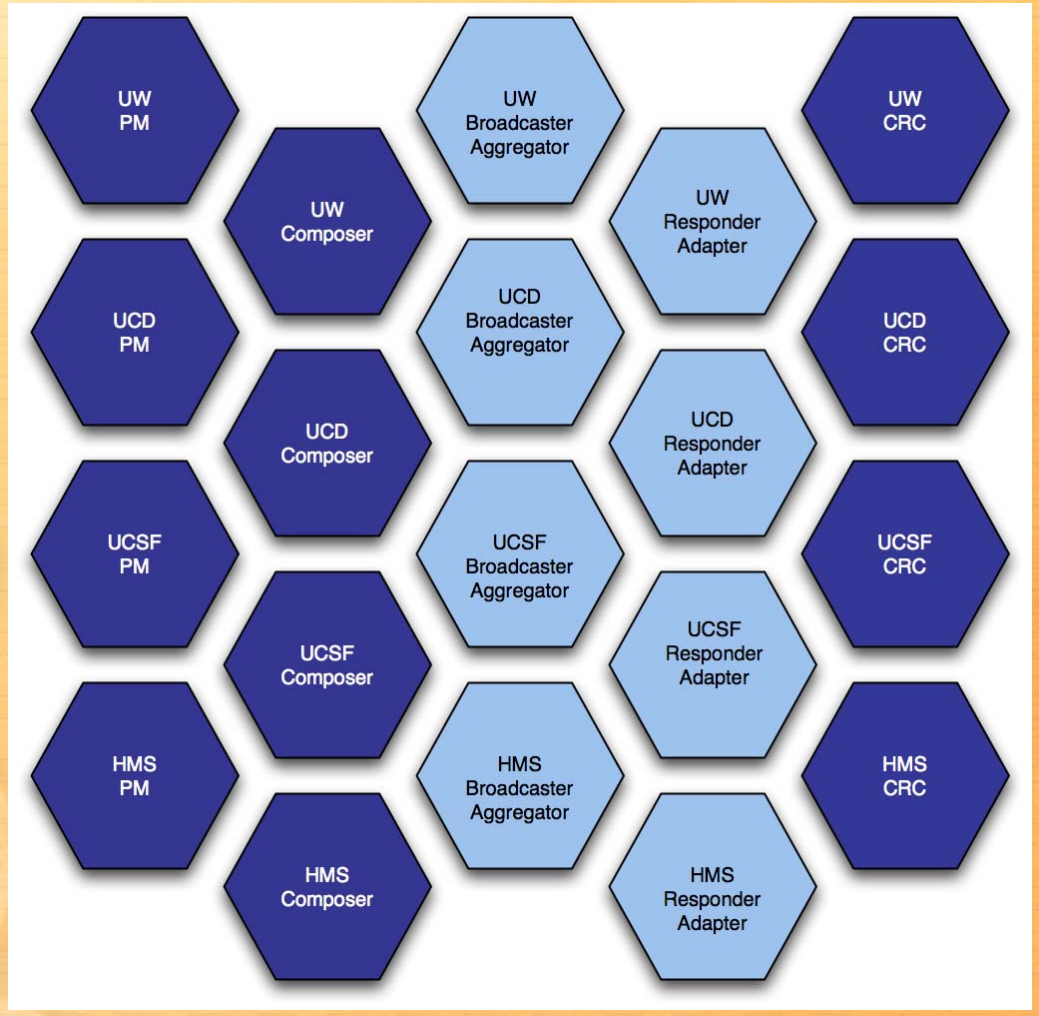
Shared Research Informatics Network (SHRINE) Distributed Queries



Central “aggregator” broadcasts query to local hospital “adaptors”, which return aggregate, “blurred” counts only

(Murphy 2009)

I2b2 CICTR SHRINE Network View

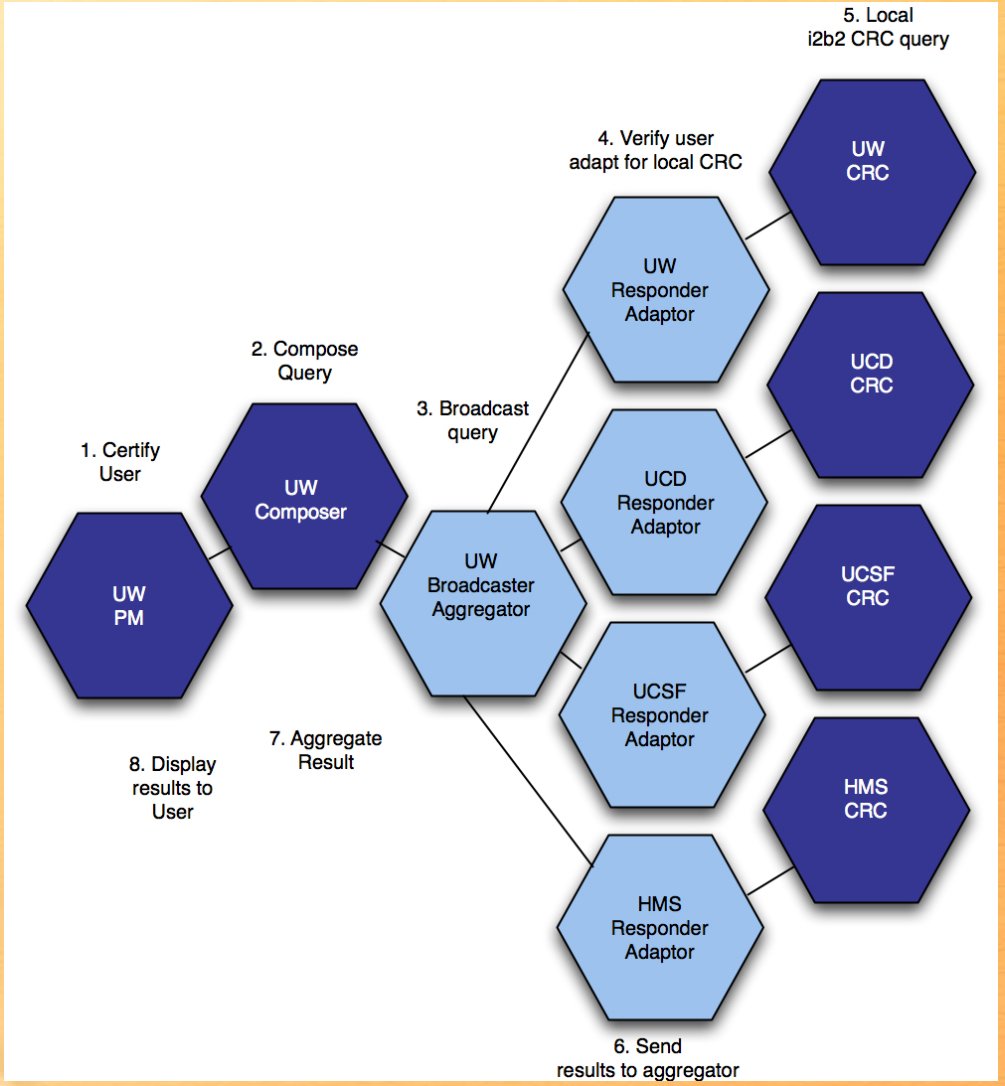


ITHS
Institute of Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics

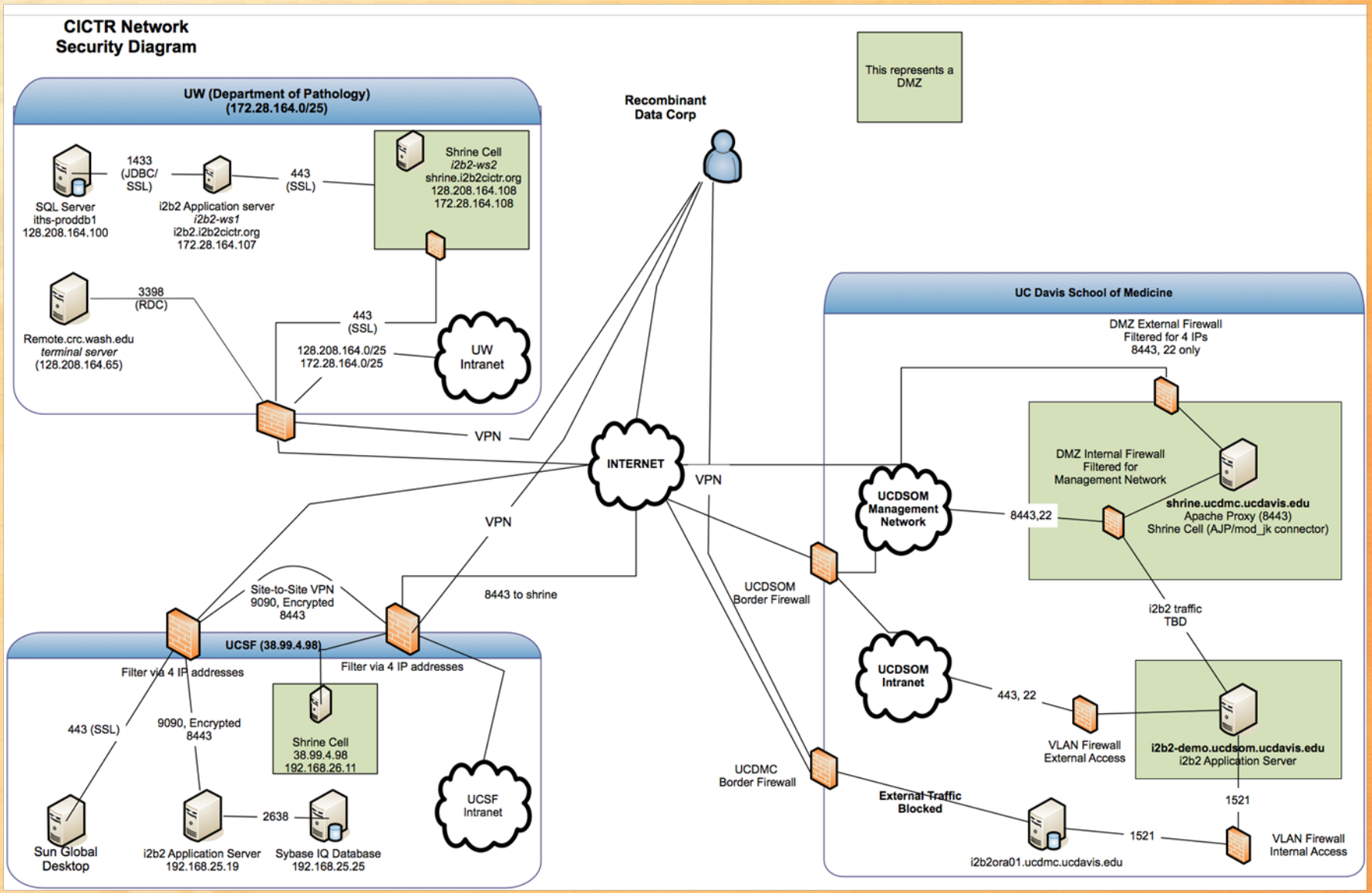
SHRINE marshalling a federated query



ITHS
Institute of Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics



ITHS
Institute of Translational Health Sciences

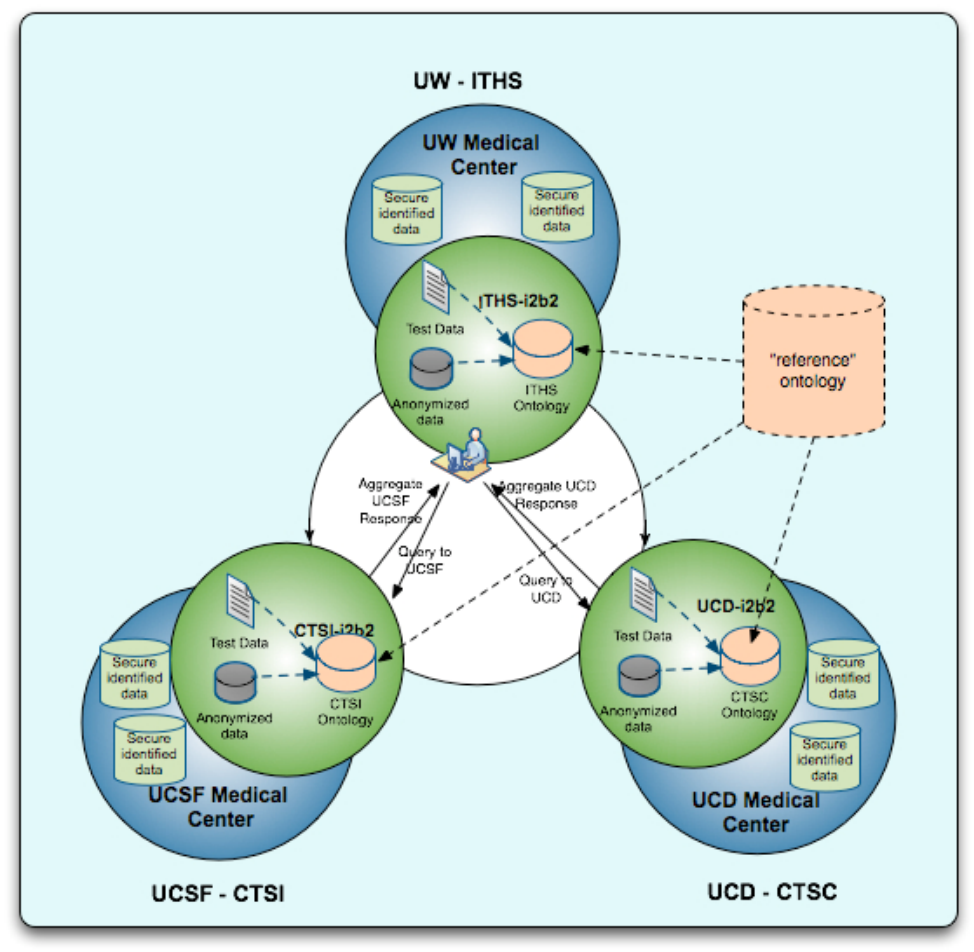
Information Exchange Pilot Project

bhi
biomedical and health informatics

Accomplishments

- Deployed secure environments with real data at 4 sites (three real partners and Harvard)
- Approaching 2.5 million de-identified patient records in three site secure network
- Ability to simultaneously query on demographics and disease diagnoses (ICD-9)
- Secondary research into:
  - usability of federated query systems,
  - anonymization approaches
  - Standards development

Phase 3: Moving to support anonymized semantically rich data discovery



- Disease/domain focus
  - Diabetes
  - Cardiovascular disease
- Pilot ability to search for "poorly characterized" disease criteria across geographically and culturally unique medical centers
- Support rare disease hypothesis generation/pruning

ITHS
Institute *of* Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics

Current I2b2 CICTR Data Elements

Available and anticipated data elements			
Demographics	**Diagnoses**	***Medications***	***Laboratory***
Age	Date of diagnoses	*Date of encounter*	*Date of lab*
Gender	ICD-9 numeric codes	*Medication name (generic/brand)*	*Lab values*
Race/Ethnicity	iCD-9 supplemental classifications influencing health status	*Dose form*	
Geocode (3 digit zip prefix)	ICD-9 supplemental classification of external causes of injury and poisoning		
Vital status			
Marital status			
Language *(tbd)*			
Religion *(tbd)*			

ITHS
Institute of Translational Health Sciences

Information Exchange Pilot Project

bhi
biomedical and health informatics

Challenges and Next Steps

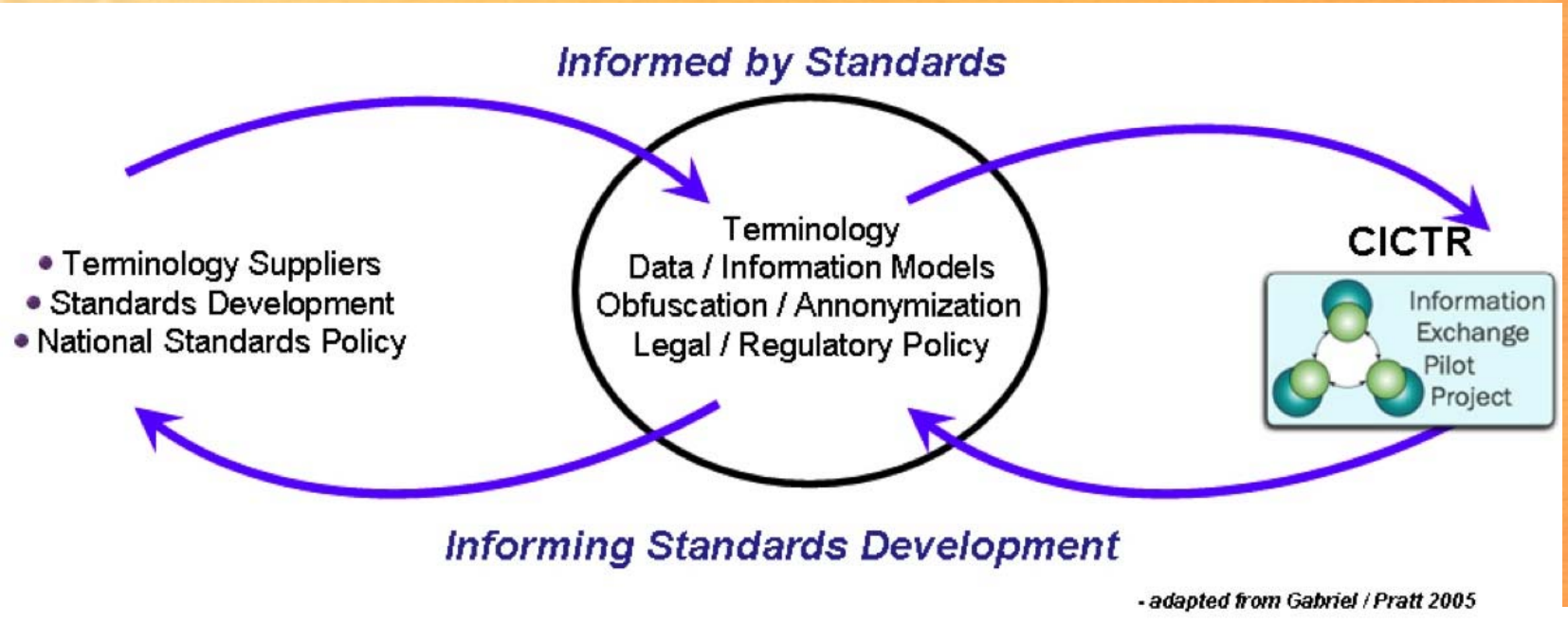
- Define/evaluate complex mapping methodologies (medications and laboratory values)
- Evaluating quality of knowledge mappings locally and network-wide, both qualitatively and quantitatively
- Develop/refine well-governed access to systems

Complex mapping challenges

- Defining common standard-based data exchange formats by LCD method
- Building or adapting tools that support federated querying/SHRINE/i2b2 environment
- Automating/enhancing labor intensive processes (ETL/Anonymization->i2b2 schemas)

Evolving Best Practices

- Developing two-way dialog with National Standards projects, organizations, development
- Increasing ability to vet evolving standards in practical research environment



Use of Standards to date

- Factors considered in the selection
  - HITSP recommended
  - SHRIMP / Harvard Ontology selections
  - Common data availability (Diagnoses) across sites
  - Widespread use: HL7 demographic value sets

Use of Standards

- Selected Standards
  - Gender: HL7 001
  - Race / Ethnicity: OMB5
  - Language: ISO 693.2
  - Marital Status: HL7 002
  - Vital Status: HL7 Entity.LivingSubject.deceasedInd
  - Religion: HL7 Religious Affiliation value set
  - Diagnoses: ICD-9 numeric, V and E codes
  - Medications: RxNorm (Ingredient table)
  - Laboratory: LOINC

Multi-site issues Informing Standards

- RxNorm feedback
- HITSP EHR-to-CTMS Value Case
  - Deals with only point to point intra-institutional data sharing
  - Participated in public comment with many to many points in mind
- IHE Redaction Services Functional Profile
- HL7 CIC Diabetes Domain Analysis Model
- End-user roles (PI's, study coordinators)

Federated mapping approaches

- Option 1: “Everybody” agree on the same target reference terminologies, and conforms
  - Seems to works for simple cases (eg. ICD-9)
  - I2b2 has tool that begins to do this (SHRIMP)

  ..but
  - Potential loss of information for complex mappings (e.g. “rich” local to LCD network to target – everybody loses)
  - What if local needs require common targets to change- who mediates?

Federated mapping approaches

- Option 2: Use a public terminology repository (e.g. OpenMDR), build local maps against public maps that are then checked-in/out/updated by services that require them (such as SHRINE)
  - caGRID world likes this (and potentially provides a bridge to caGRID environments)
  - Current tool at UCSF (Ontomapper) is doing this
  - Not much on the “agnostic” repositories yet (needs content)
  - Incompatible with SHRINE at present..

Federated mapping approaches

- Option 3: Blend/test/deploy/test these
- Leverage SHRINE/SHRIMP/I2b2 terminology philosophy
  - Extend SHRIMP to use OntoMapper capabilities
  - Let the site decide
  - Will require additional SHRINE/i2b2 configuration/development
- … Will report results soon…

Evaluating end-user needs

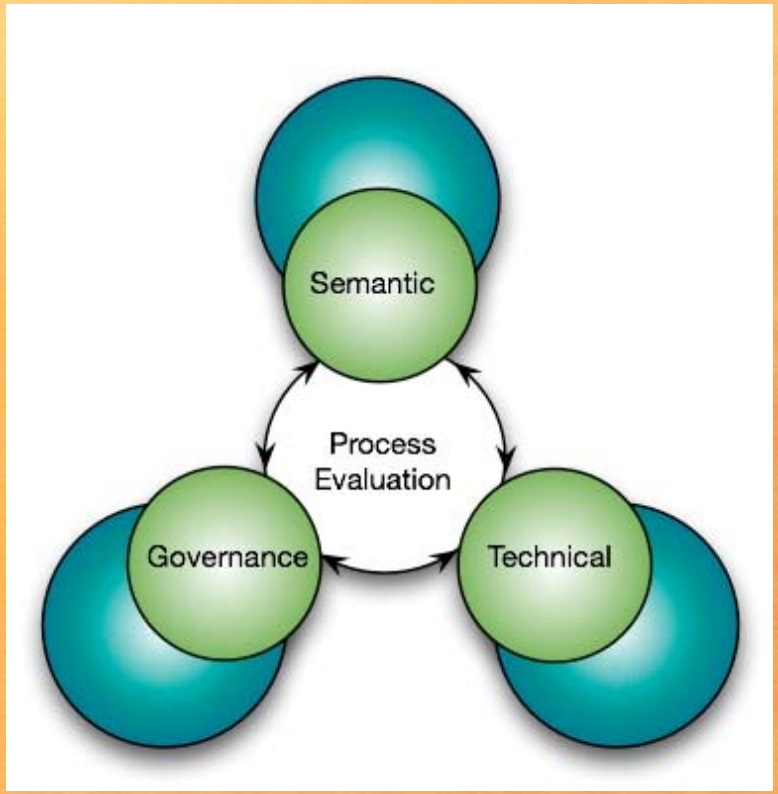
- Resulting data organization may not intuitively support how researchers create structured queries
  - Testing use of system expectations locally and nationally via Davis evaluation group
  - Planning near-term focused diabetes group

Providing access for researchers

- Significant institutional sensitivity to the use of such systems
- Hypothesis: best researcher use is a governed approach that puts them in the query-seat
  - How to facilitate this? Currently data is technically “not human subjects”, yet sensitivity and emphasis on secure control remains

Pandora's box issues

- There is a risk of being successful…
  - Business intelligence
  - Setting wrong expectations/freaking people out
  - Create unnecessary new branches of code/process
- Best practices/technologies are evolving..
- Maintaining scope and control
  - Protecting the patients
  - Protecting the partners
  - Protecting the developer



Acknowledgements

- John Gennari (Co-I)
- Fred Wolf (Co-I)
- Peter Tarczy-Hornoch (Co-I)
- Justin Prosser (Sysadmin)
- Elishema Fishman (RA)
- Parmit Chilana (former RA)
- Marcin Porwitz (rotation)

•Michael Kamerick (UCSF)
•Rob Wynden (UCSF)
•Julie Rainwater (UCD)
•Kent Anderson (UCD)
•Este Gehaghty (UCD)
•Davera Gabriel (UCD)
•Andy McMurray (Harvard)
•Aaron Abend (Recombinant Data)
•Aaron Mandel (Recombinant Data)
•Zach Kohane (Harvard)