Translational Research Information Systems:

Building the Integrated Data Repository

Michael Kamerick
Director, Academic Research Systems
co-Director, Biomedical Informatics,
Clinical and Translational Sciences Institute
University of California San Francisco
21 January 2009

UCSF University of California San Francisco

CTSA Clinical & Translational™ Science Awards

CTSI

Outline

- Definition
- Section I – The Work of Research
  - The Value Proposition – Why build an IDR?
  - Value to current research methods
  - New methods made possible
  - Social and Regulatory, i.e., Governance Issues
  - CTSA activities
- Section II – The technology
  - Technical Governance
  - Data Sharing
  - UCSF
- Summary

Integrated Data Repository Definition

From the CTSA Data Repository Interest Group Wiki:

We define an Integrated Data Repository as a very large-scale database containing data from the full array of systems in a biomedical enterprise, including clinical systems, life sciences (genomics/proteomics), research, billing, registries, clinical trial systems, and more. The purpose of an IDR is to support a wide range of activities within the biomedical research enterprise, including but not limited to hypothesis testing, cohort development, genome/phenome matching, genome-wide association studies(GWAS), development of quality measures, and general population based studies.

The Value Proposition

- Taking time out of the research cycle
  - 17 years from discovery to practice!
  - Manually intensive methods of data collection
  - Outdated modes of dissemination
  - Much faster cohort selection, the #1 use case
- Recast funding dollars
  - Services, not capital or salary
- Create/Enable new research models

Typical Research Query

- I was wondering if there was a mechanism in place for UCSF to do retrospective patient analyses using icd-9 code searches/discharge diagnoses. For example, we were interested in looking at our patient series of children <21yo with heparin induced thrombocytopenia in the last 5 years. Is such a query available?

The Current, Painful Response

- No
- Comprehensive response will require data from up to 8 systems, some of which are still on paper!
- Different system owners, most not helpful.
- HIMS (Paper Chart), MAR (paper), UCare (newer, EMR), TSI(Billing), WorX(Pharmacy), Pixis(Cart Dispensing),PICIS(Peri-operative), STOR (Older EMR).
- How long? 1 year if lucky? 2 years? Never?

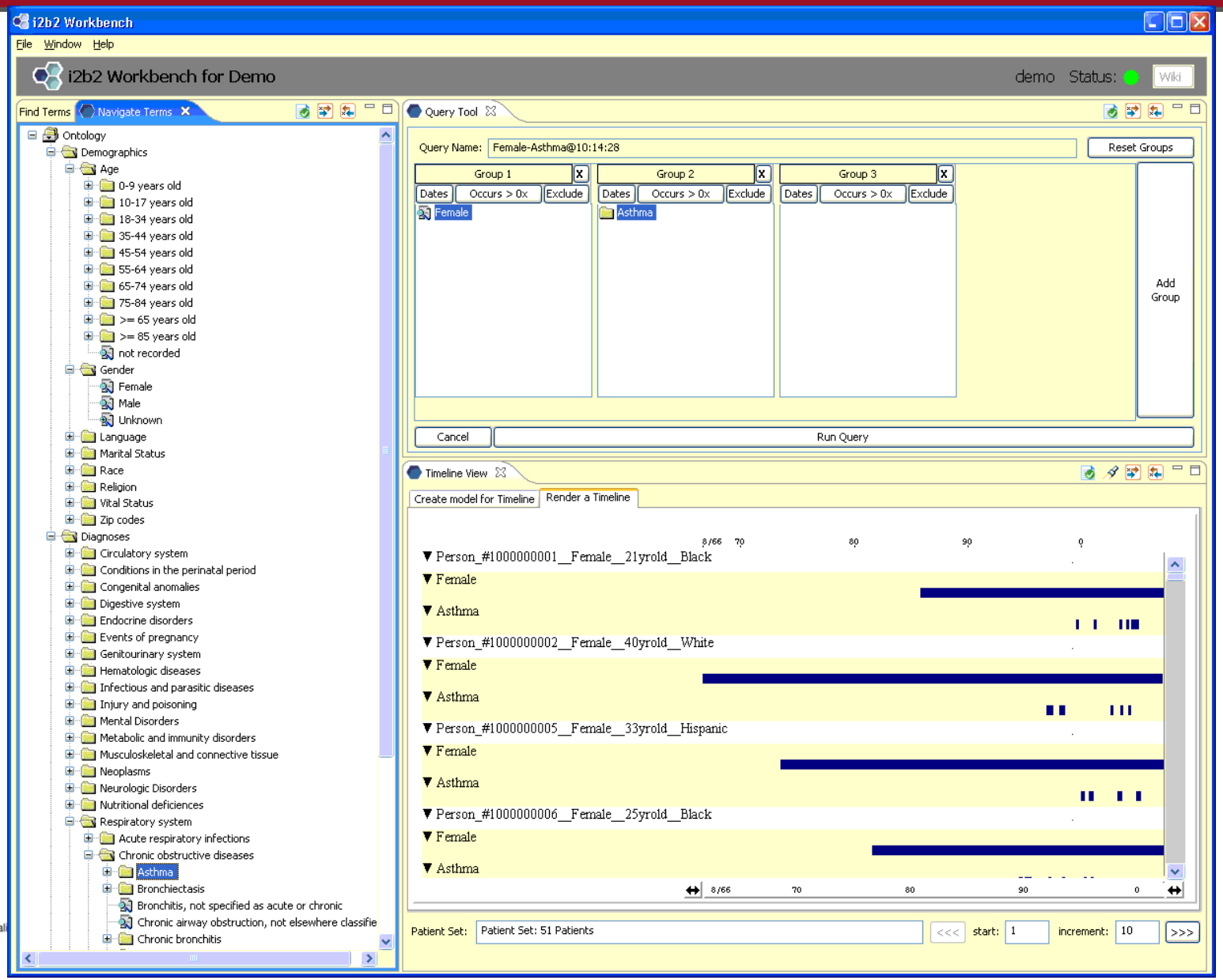
The Current Painful Methods of Data Gathering

- Intensively Manual
- Review of paper charts
  - 3 years for flu study of studies
  - Exposes all individual data to investigator
- Manual screen scraping
  - Study coordinators transcribe records from EMR into spreadsheets.
  - Time consuming, error prone,
  - Zero security.

Shortening the Cycle

- Three years becomes 3 weeks, 3 days, 3 hours, 3 minutes.
- Information is managed in secure, professional environments
- Proxy chart review
- i2b2 Workbench as example

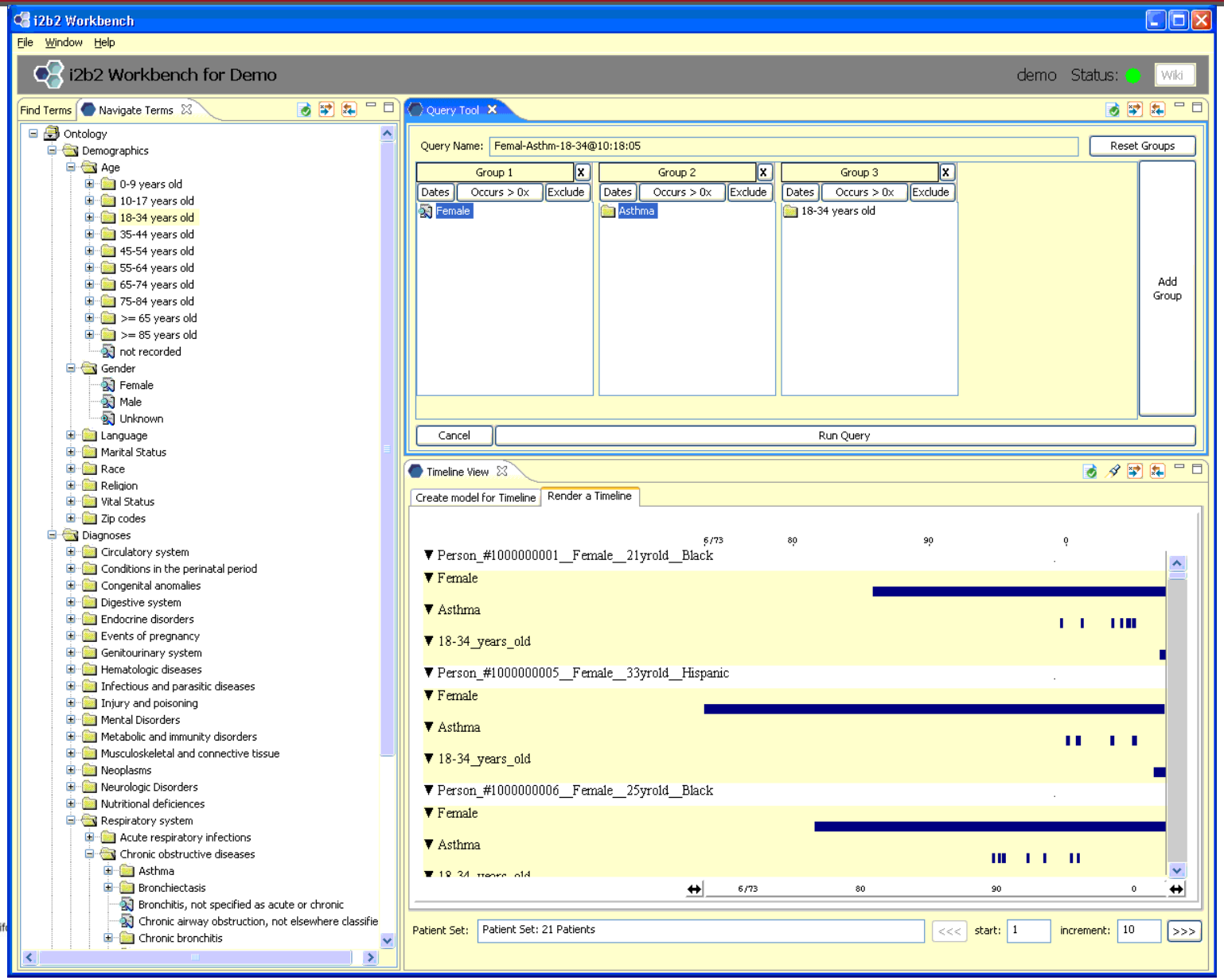
i2b2 Workbench Example 1



UCSF University of California San Francisco

CTSI

i2b2 Workbench Example 2



UCSF University of California San Francisco

CTSI

Recasting Funding Dollars

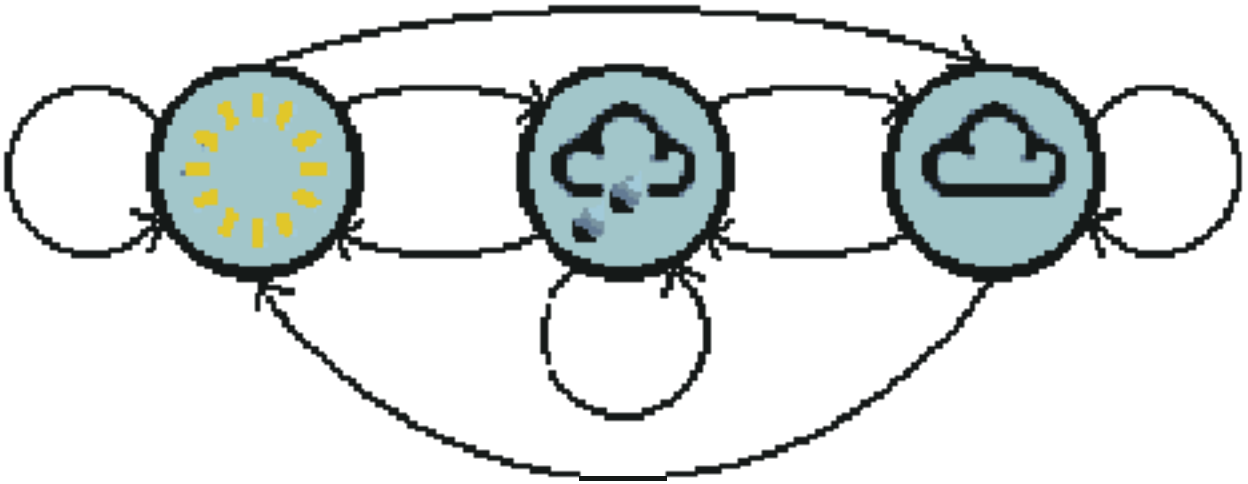
- Centralized Research IT Infrastructure vs. a thousand MS Access db's.
- Enterprise scale, managed, secure
- Control point for release of PHI
- Service/subscription vs. capital/payroll

New Research Paradigms

- Ocean of Data
  - Ventner, Wired article
  - Kohane diabetes analysis
  - Neurocommons/Science Commons project
  - Delineate large effects in small populations and small effects in large populations.
- Virtualized Clinical Trial
  - Mark Weiner's work

Enables multi-disciplinary collaboration

- Disease modeled as state machine vs. Markov model
  - require enormous amounts of data to be deterministic.



The IDR is a Disruptive Technology

- Changes the way biomedical research is done
- Changes the speed of research
- Raises new possibilities
  - Statistical methods vs. RCT
- Increases security and access simultaneously
  - Proxy chart review
  - Single control point for release of clinical data

The Necessity of Automation

- Productivity gains of the last 30 years predicated on automation
- The Information Economy - Fedex, Wal-Mart, Google
- Research IS an information economy
  - The value of a tissue bank is ultimately the information that can be derived from analysis of the samples
  - Managing that information becomes as important as managing the samples.
  - Tissues may be a scarce resource, but information about those tissues can be reproduced at almost no cost.

- Many technological problems solved in other industries
  - Healthcare and research lag behind in application and investment
  - Great advances could be made using today's technology
- However…

The Challenge of Narrative Text

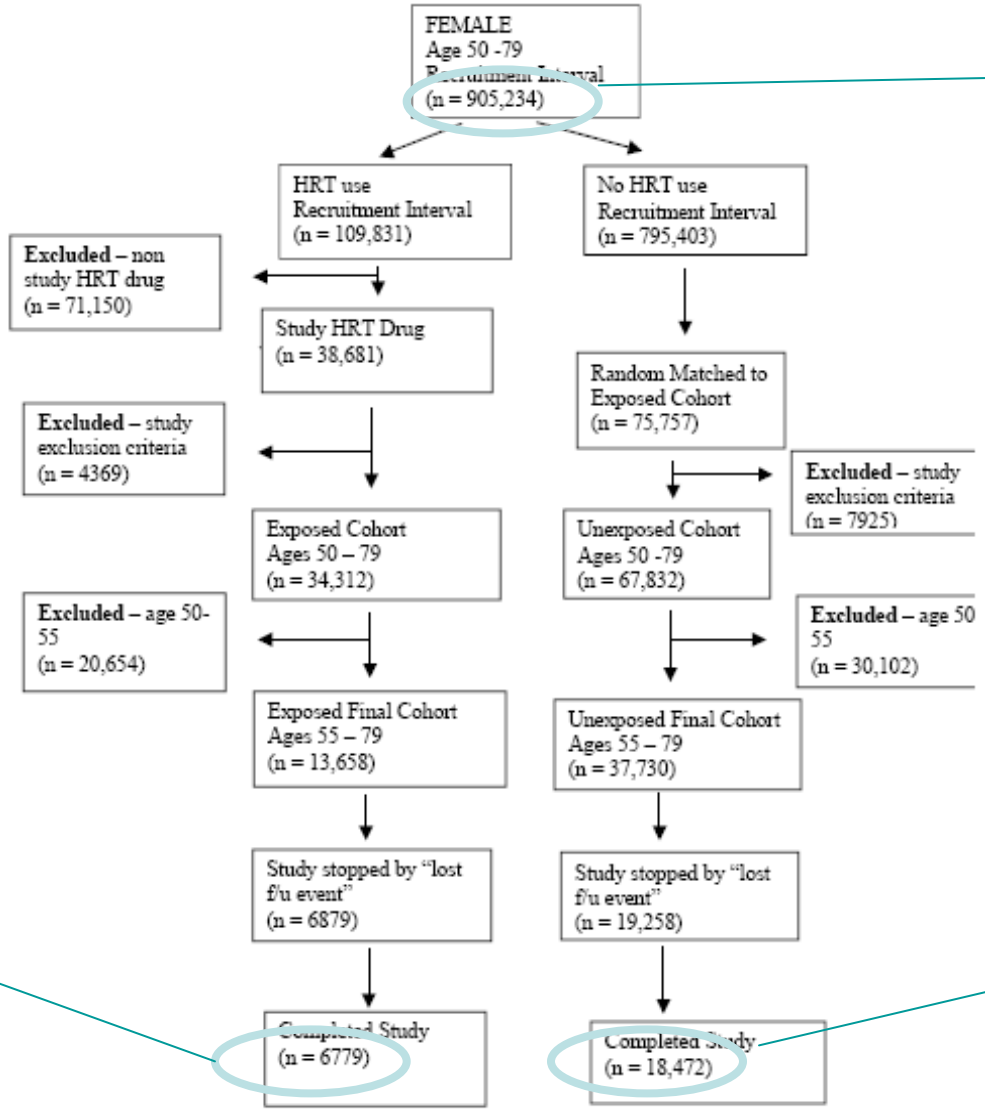
- Automation requires computable data
  - Dominance of narrative text in healthcare
  - Word vs. Excel
  - Natural Language Processing (NLP)
    - Best solutions typically get only 70% accuracy
    - UPMC claiming much better rates
    - CTSA has begun NLP interest group, led by Zak Kohane

Secondary Use of Healthcare Data

- Predominance of narrative text (see above)
- Data Quality is the other big issue
  - Always worse than RCT data
  - Precise data not always required for care decisions
  - Large data sets needed to mitigate lower quality of data
    - ref. Mark Weiner's work.

Subject Selection

(aka why you need to start with a large database)



905,234

6,779

18,472

UCSF University of California San Francisco

CTSI

Governance Examples

- Oversight committees
  - Faculty boards, Privacy Office, ISO
- Documents
  - IRB protocols, MOUs, BAA, Certificates of Confidentiality
- Patient's Rights
  - Opt-out vs. Opt-in?
  - No Opt-out?
    - Stanford, Partners
  - Challenging Opt-out
    - UCSF
  - Clear Opt-out
    - Vanderbilt
  - Special Cases – Prisoners, VIPs, Opt-outs

Examples, continued…

- Data Ownership questions
  - Clinician/Investigator vs. Institutional
- Stakeholders
  - Hospital IT, IRB, Privacy Office, Security Office, Medical Records, Legal Office,
- Security requirements
  - AuthN/AuthZ, Two Factor AuthN, Local disk encryption, Securely managed storage
- Limited Data Sets, Honest Broker function
- Small Cell Results

Interaction With IT Governance

- IDR within Hospital IT organization
  - Mayo, UPMC, St. Jude's
  - Much less institutional conflict
  - IDR project likely to rank lower in priority schemes than more urgent hospital projects
  - May be much harder to add in non-hospital data sources
- IDR in IT organization separate from Hospital IT
  - Stanford
  - Long, hard road to intra-institutional agreements
  - IDR project can be prioritized independently of Hospital IT
  - Easier to include non-hospital data sources
- Federated IDR - crosses IT organization boundaries
  - UCSF
  - Architecture maps to stakeholder boundaries
  - Best or Worst of both worlds?

IDR Regulatory Environment

- Extremely challenging and complex
- Goes well beyond HIPAA
- Contradictory
  - May not be possible to be compliant
  - Laws written without regard to consequences
- IRB policies may be outdated and insufficient
  - IT staff burdened with policy decisions
- Very difficult to provide sufficient utility to researchers while fully protecting patient privacy
- IDR use can be especially sensitive
  - Patients generally NOT explicitly consented

Federal Laws and Regulations

- **HIPAA**
  - Health Insurance Portability and Accountability Act
- **FISMA**
  - Federal Information Security Management Act
- **FERPA**
  - Family Education Rights and Privacy Act
- **GINA**
  - Genetic Information Non-Discrimination Act
- **21 CFR Part 11**
  - Code of Federal Regulations Electronic Signature
- **Sarbanes Oxley**
- **NIST 800-53**
  - National Institute of Standards
- **E-Discovery**
  - Federal law for preserving and protecting electronic data in Federal civil lawsuits.
- **NIH Certificate of Confidentiality**
  - Protection against E-Discovery
- **FIPS 140-2, 196, 199, 200**
  - Federal Information Processing Standard

State and Institutional Laws and Regulations

- State of CA
  - Title 22
    - Definition of the Medical Record
  - SB 1386
    - Notification Requirements
  - AB 1298
    - Extension of 1386 to include “Medical Data”
  - SB541, AB211
    - Specify penalties for individuals and institutions for “negligent” handling of medical data.
    - Up to $250,000
- UCSF/UC
  - 650-16
  - ECP
  - UCOP IS2 and IS3

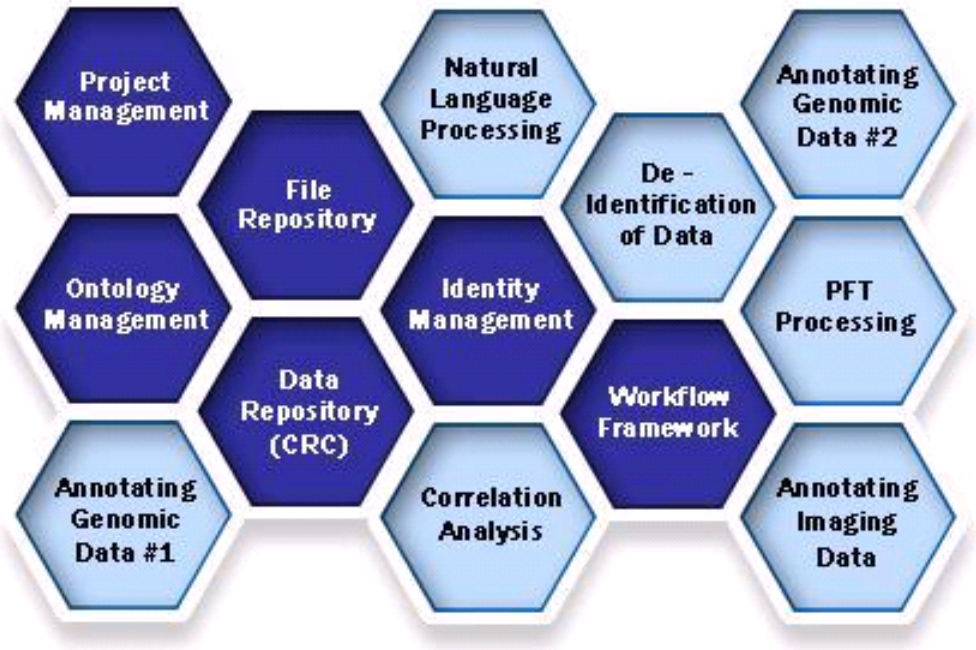
CTSA

- IKFC - Informatics Key Function Committee
  - Loose affiliations
  - No data coordinating center
  - No IT standards
- Multiple Interest Groups, Projects
  - Data Repositories, Data Sharing, Education, Standards and Interoperability, Inventory, Human Studies DB, Collaboration Facilitation, National Recruitment Registry, others.
- Data sharing
  - CICTR(UW, UCD, UCSF)

Data Repository Interest Group Activities

- **Ontology Mapping Service**
- **Integration of i2b2 with caGRID**
- **Data Sharing Across Repositories**
- **Best Practices Symposium**
- **Repository Inventory Survey**
- **Governance Documents**
- **Conference Calls**
- **Integration of Molecular and Clinical data**
- **EMPI**

The i2b2 Hive



Technical Data Governance

- Classic Data Warehouse Design
  - Inmon, others.
  - Enterprise Data Model
  - All data transforms and encodings done up front, during ETL
  - Long negotiations between stakeholders to get agreement on the model.
- Late Binding Design
  - Minimal ETL.
  - Customized data models based on user preferences and beliefs
    - Supports multiple terminologies/ontologies
    - CTSA Ontology Mapper
  - Diverse data models expressed as views or physical marts

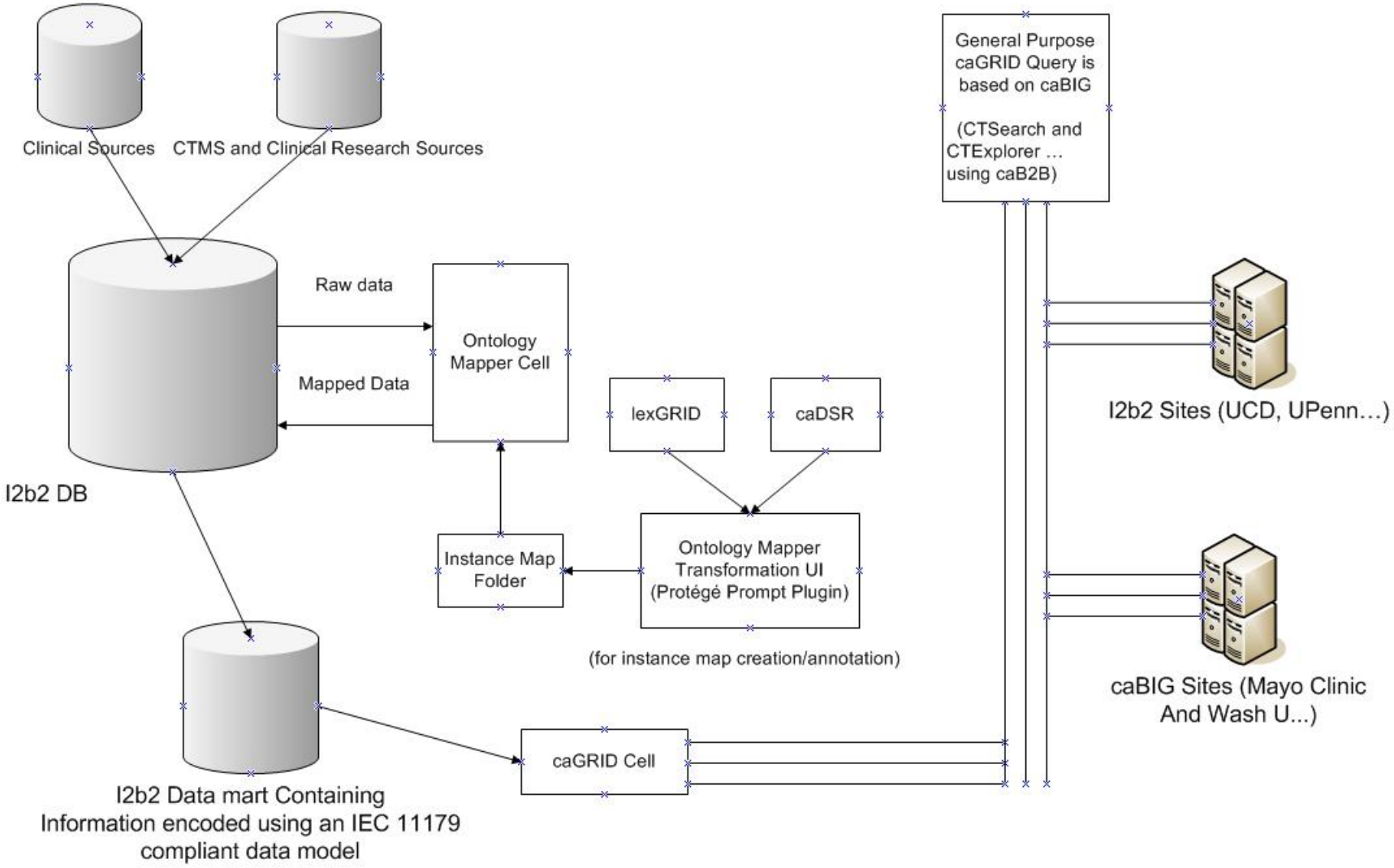
Ontology Mapper

- Written as an i2b2 cell
  - General purpose instance mapper
  - Translates messy local data into one or more standard formats
  - Maps local data into Ontologies
- Maps will be created and annotated in a Protégé Prompt plug-in and can be shared over HL7 CTS II both as open source or as commercially sold assets
- Maps contain routing, provenance information and a scriptlet payload of SQL, Perl, SparQL, Horn or R
- The Ontology Mapper Cell within i2b2 is a collaborative effort involving UCSF, UCD, Rochester, UPenn, and U Washington
- This has been a highly active collaborative effort which is now in an Alpha release cycle

i2b2 caGRID Cell

- The caGRID Cell is a development project which is a collaboration of OSU (Ohio State) and UCSF
- This component allows any i2b2 data mart, which has been translated into standard format by the Ontology Mapper, to share data over caGRID
- This system will allow i2b2 to share data (a federated query) across any caGRID based data source (not just between other i2b2 instances)

CTRgrid Design



CTRgrid Components

- NCI caGRID
  - Well defined grid for sharing data in a secure and semantically complete manner
  - Designed for cancer, but the NCI wants to generalize it
- NCBC i2b2
  - The software platform for the Integrated Data Repository
- CTSA Ontology Mapper
  - Takes the raw data of the repository and turns it into a structured, study domain specific model that can be shared across caGRID
  - First CTSA developed software
  - Led by UCSF
  - Incorporated into HL7 CTS II standard

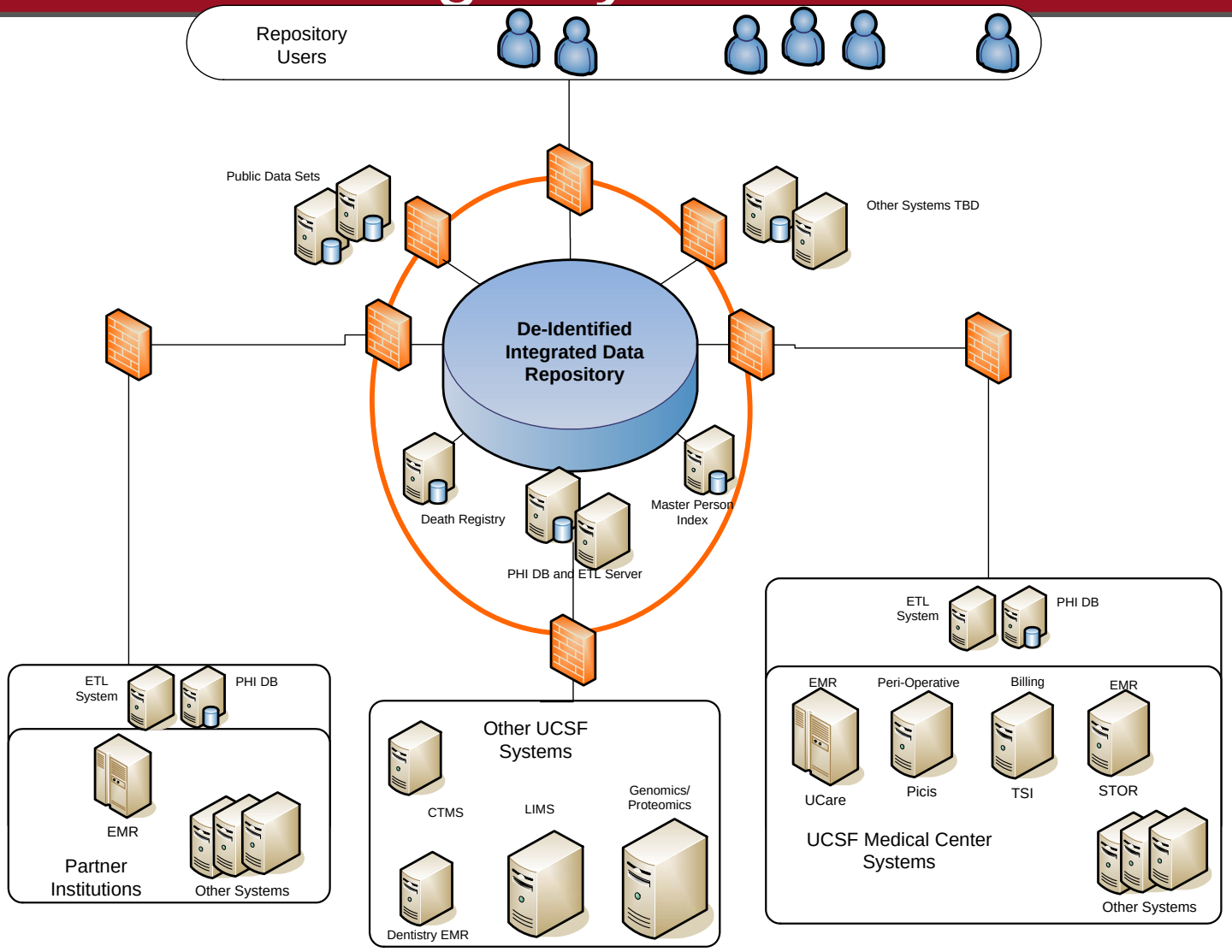
Near Term Projects

- Human Studies DataBase - Ida Sim
  - UCSF, Mayo, Wash. U
- CHORI (Dentistry) – Joel White
  - UCSF, Harvard, Tufts, UT Houston
- STIRS (Radiology) – Max Wintermark
  - UCLA, Georgetown, Wash. U, Edinburgh, Nottingham
- Pediatrics Rare Disease – Jennifer Puck
  - UCSF, UT Houston, Harvard, Duke, Emery, OHSU, Vanderbilt, Chicago, Hopkins, Columbia
- Quality Network – Andy Auerbach
  - Northwestern, Tufts
- CTSA i2b2 Adoption – Russ Cucina
  - U. Wash, UCSF, UC Davis

UCSF Activities

- i2b2, Sybase IQ integration
- MyResearch Portal
  - Remote desktop for managing research data
- Virtualized server infrastructure
- Managed Services vi ARCAMIS/ITN
- Service Model of Research IT
- CTRgrid
- General Security Model
- Workflow Models
- Governance difficulties
- Public data sets

Integrated Data Repository: Design by Governance

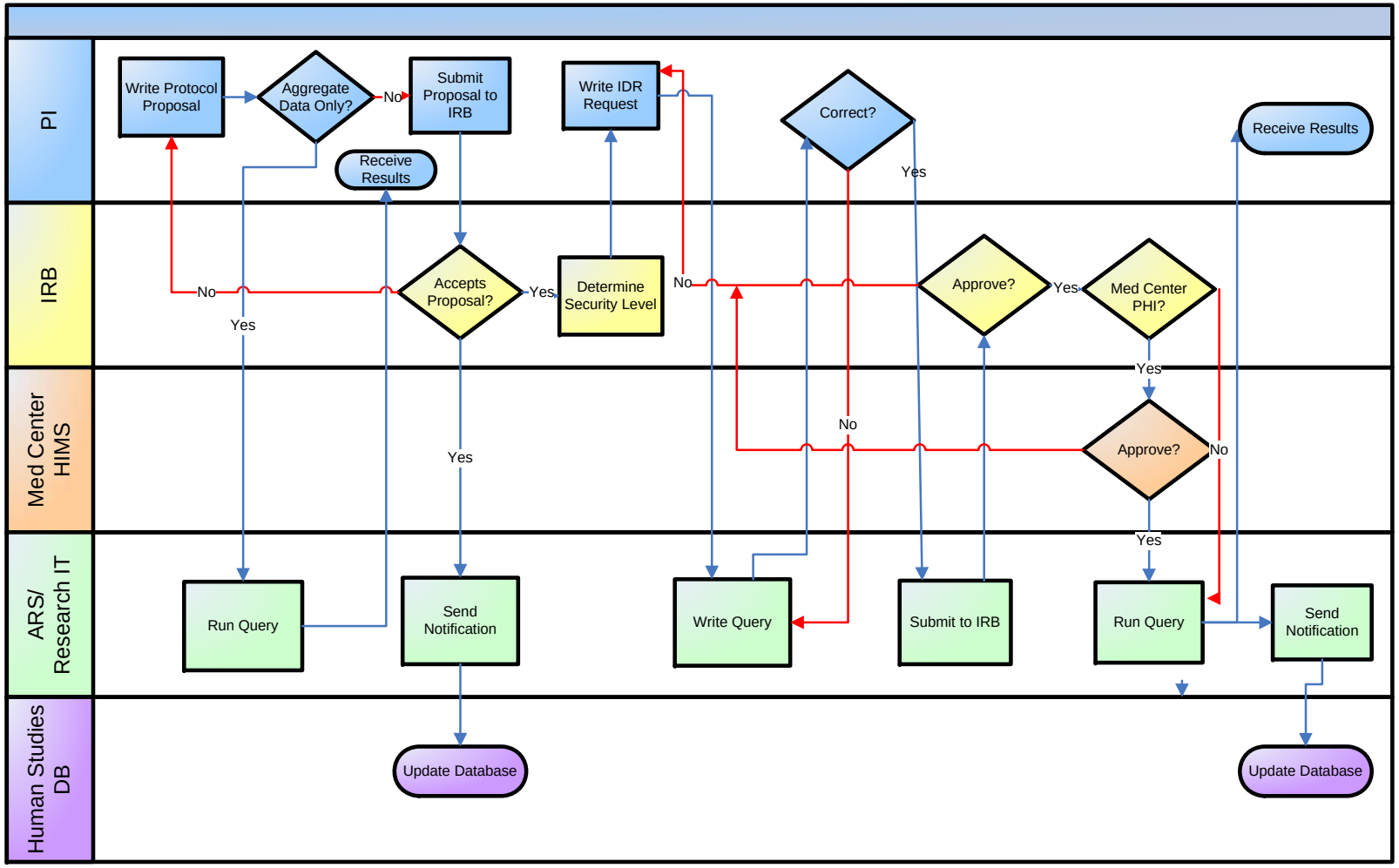


UCSF University of California San Francisco

CTSA Clinical & Translational Science Awards

CTSI

Research Data Request Workflow

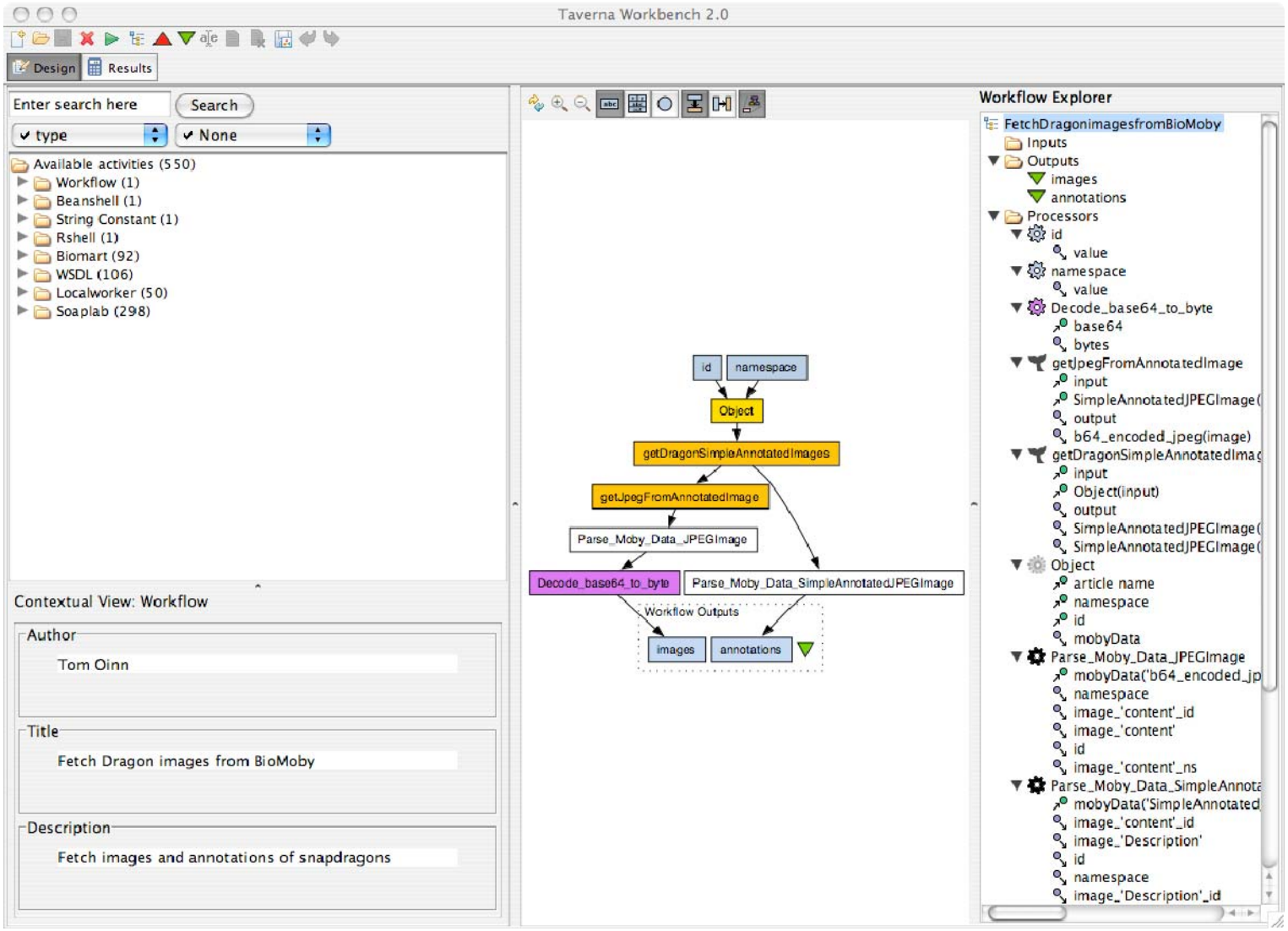


UCSF University of California San Francisco

CTSA Clinical & Translational Science Awards™

CTSI

Taverna Scientific Workflow



UCSF University of California San Francisco

CTSA Clinical & Translational Science Awards

CTSI

Summary

Building Integrated Data Repositories presents exciting challenges and wonderful opportunities for advancing the frontiers of biomedical research.

Many institutions across the country, inside and outside of the CTSA are engaging in this task.

Networking these repositories will be the way we arrive at the “Ocean of Data”.

Shortening the research cycle shortens the translational cycle from discovery to clinical practice, advancing health worldwide.

To the Future, the Undiscovered Country