Electronic Health Records with Cleveland Clinic and Oracle Semantic Technologies

David Booth, Ph.D., Cleveland Clinic (contractor)

Oracle OpenWorld

20-Sep-2010

Latest version of these slides:

http://dbooth.org/2010/oow/

Outline

- Background on SemanticDB project
- Current state of electronic health data
- Cleveland Clinic semantic initiative and strategies
- Cleveland Clinic experiences implementing this initiative

Cleveland Clinic's Heart and Vascular Institute

- **Patient care:**
  - **Ranked [#1 in heart care] by *US News and World Report* for the past 16 years**
  - **Over [4,000 cardiac surgeries] performed in 2009**
- **Research:**
  - **~130 journal articles/year**

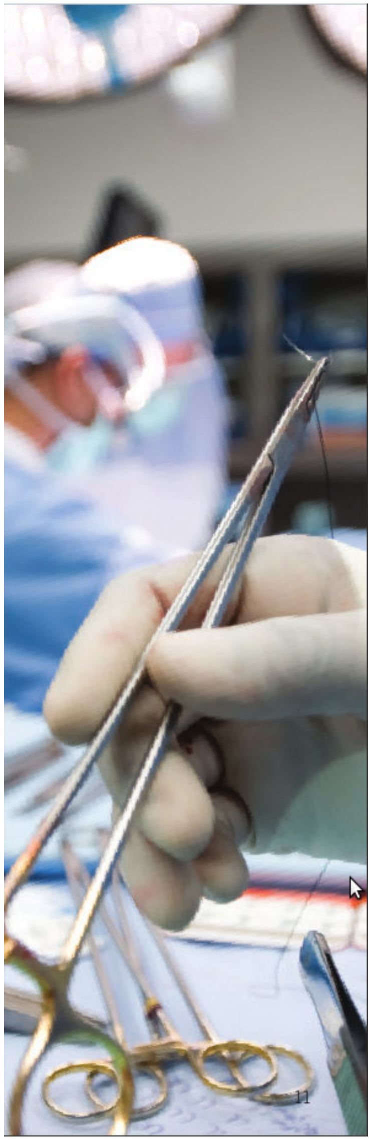
Cleveland Clinic SemanticDB Initiative

TimeLine:

- **1997-2002 Small proof of concept studies**
- **2003 Launched development project**
- **2004 Created Patient Record ontology**
  - **(>4000 classes & > 400 relations)**
- **2007 Began Cycorp collaboration**
- **2007 Converted 200K patient's data to RDF (~120 million triples)**
- **2008 Live production system released**
- **2010 Move to commercial semantic platform**

SemanticDB Project

- **Project in Cleveland Clinic's Heart and Vascular Institute**
- **Applies semantic web technology to support data needs for:**
  - **Research**
  - **Quality reporting (i.e., measuring quality of care)**



Patient-centric vs. population-centric data views

- **Patient centric:**
  - **Optimized for individual patient treatment**
  - **Used by care givers**
- **Population centric:** **Our focus in this talk**
  - **Optimized to look across many patients**
  - **Used for outcomes research & measuring quality of care**
  - **E.g., which treatments produced the best outcomes?**

Semantic web technology

- **RDF: Data model framework**
  - **W3C standard**
  - **Permits very flexible information capture as *<subject, property, object>* triples, e.g.:**

    _:bloodPressure231 :diastolicMPa 80 .
- **OWL: Ontology language (more on this in a moment)**
  - **Used to define specialized ontologies**
- **SPARQL: Query language for RDF**
  - **A little bit like SQL**

What is an ontology?

- **Set of <u>concepts</u> and the <u>relationships</u> between them**
- **Analogous to a database schema, but:**
  - **Attempts to capture semantics (i.e., meaning) rather than structure of data**
- **Used to support machine processing (inferencing)**
- **E.g.**
  - **:AorticValve :physical-part-of :Heart .**
  - **:AorticValveProcedure :subClassOf :HeartProcedure .**

Patient Record Ontology

- **Captures patient record concepts and relationships:**
  - **Patient information (e.g., demographics)**
  - **Patient events (e.g., surgeries performed, medications taken)**
- **Used to classify medical procedures, events, medications, etc.**
- **Supports inferencing, e.g.,**
  { :x a :AorticValveProcedure . }
  ==> { :x a :HeartProcedure . }
- **Enables simpler queries, e.g.,**
  SELECT ?x
  WHERE ?x a :HeartProcedure

Why RDF and semantic web technology?

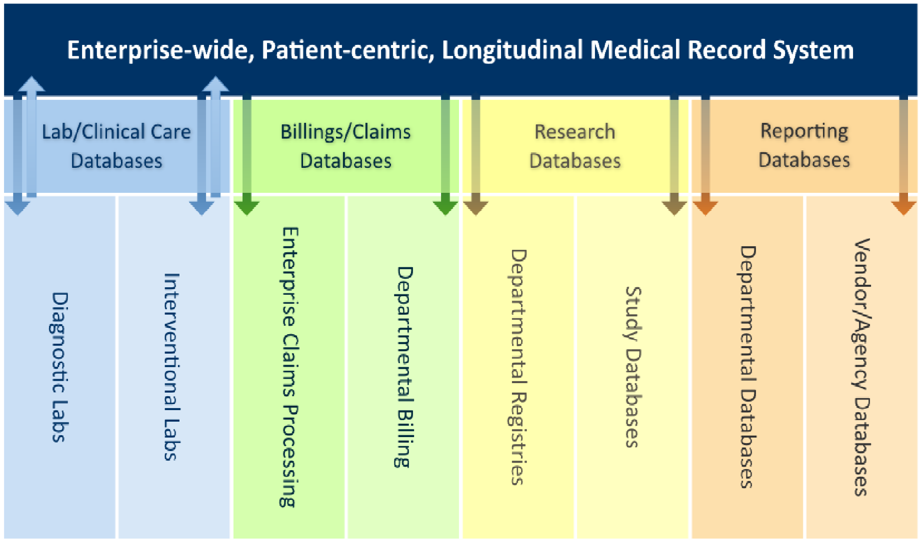
Compared with relational representations:

- **Disparate data can be integrated more easily**
- **Data can be queried more flexibly**

Current Electronic Health Data

Data Sources:

- **Enterprise Electronic Medical Records (EMRs)**
- **Lab databases**
- **Billing/Claims databases**
- **Research data registries**
- **Reporting databases**



Enterprise EMRs

A complete record of patient encounters including demographics, medical history, medications, tests, images, treatments, etc.

Benefits:

- **Comprehensive scope for enterprise**
- **Accessible to human users across the enterprise**

Challenges:

- **Mostly narrative content**
- **Structured content often inaccessible for significant periods of time, and difficult to retrieve**

Lab Databases

Patient data captured during specific medical tests and treatments including indications, methods, results, and complications.

Benefits:

- Mostly structured content amenable to use by computers

Challenges:

- Restricted scope to specific procedure
- Locally defined terms
- Limited accessibility

Billing/Claims Databases

Data collected to support billing for specific procedures and diagnoses for patients

Benefits:

- **Use of national and international standard codes and terms**
- **Structured data with enterprise-wide scope**

Challenges:

- **Terms of limited or misleading clinical relevance**
- **Can be difficult to access**

Research Data Registries

Patient data collected to support outcomes research in specific domains

Benefits:

- **Structured data**
- **Consistent, longitudinal data vetted through use in studies**

Challenges:

- **Restricted scope**
- **Locally defined terms**
- **Data silos with limited accessibility**

Reporting Databases

Patient data collected for specific reporting to regional and national quality monitoring groups

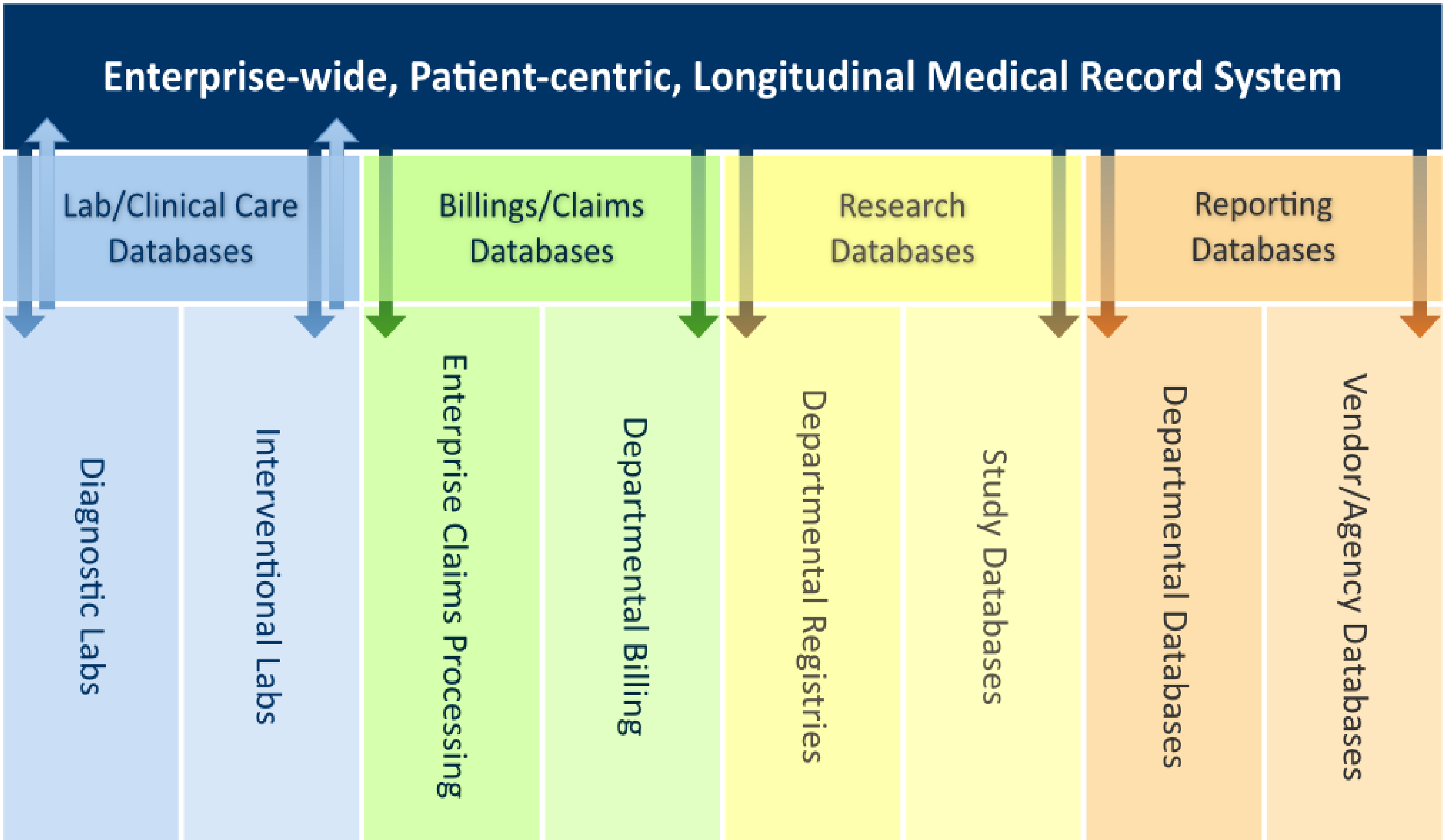
Benefits:

- **Term definitions consistent across enterprises**
- **Structured data**

Challenges:

- **Restricted scope**
- **Definitions of the same terms vary among reporting databases**

Electronic Health Data Ecosystem



How to Accomplish Meaningful Use?

- **Infrastructure needed for meaningful use:**
  - **Localized control of data collection**
  - **Centralized control of data <u>definitions</u>**
  - **Machine and human readable definitions of all data elements**
  - **Structured data amenable to machine processing**
- ***Semantic technology can help in building this infrastructure!***

Cleveland Clinic SemanticDB Initiative

Goal: Make population-centric data available and useful to clinical investigators and administrators across the enterprise to:

- **Improve reporting of health care quality metrics**
- **Facilitate clinical research (study data collection, cohort identification, analysis dataset creation, etc.)**

Cleveland Clinic SemanticDB Initiative

HOW? Reduce barriers to population-centric use of electronic medical data by:

- **Increasing data interoperability:**
  - data accessible from one system and usable by others
- **Increasing data reusability:**
  - data useful for multiple and novel purposes
- **Reducing data silos:**
  - data accessible from centralized source(s) through integration and federation
- **Reducing data redundancy:**
  - data collected once and usable by all

Cleveland Clinic SemanticDB Initiative

Strategies:

- **Build centralized/federated semantic data repository**
- **Define and collect stable core data elements and clinical facts**
- **Define RDF data models augmented by domain and upper ontologies**
- **Link RDF instance data with ontologies and rules to support inference, query, and derived views**

Cleveland Clinic SemanticDB Strategies

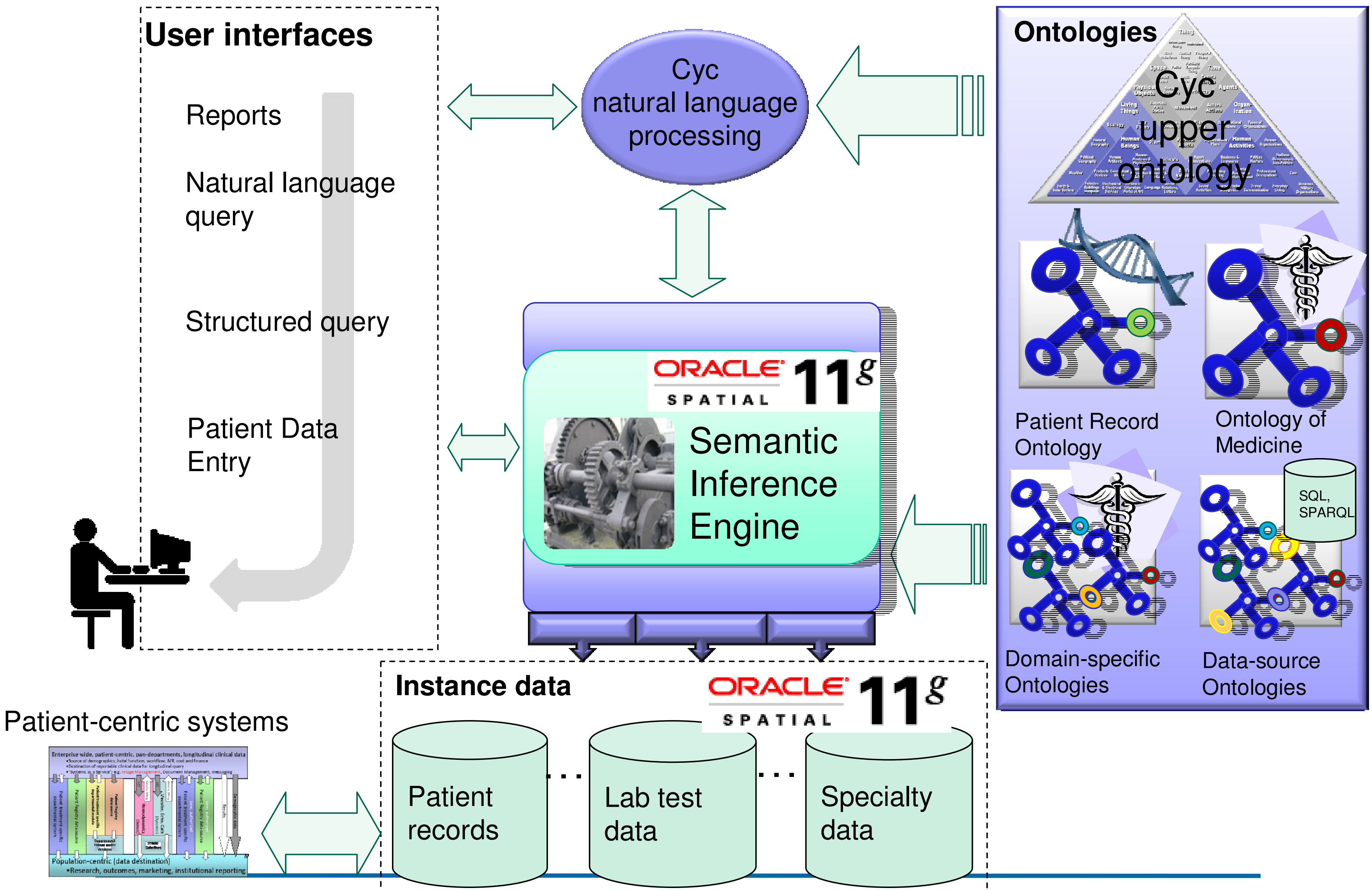
- **Build Centralized/federated semantic data repository**
- Define and collect stable core data elements and clinical facts
- Define RDF data models augmented by domain and upper ontologies
- Link RDF instance data with ontologies and rules to support inference, query, and derived views

Why a Semantic Data Repository?

Compared with ETL-based warehouse:

- **Easier data integration**
- **Removes syntactic barriers**
- **Provides robust framework for reconciling semantic discrepancies**

Cleveland Clinic SemanticDB Platform



Migration to Oracle Spatial 11g

- **In 2010 we migrated from open source Triclops RDF database to commercial RDF database**
- **Selected Oracle Spatial 11g**
- **Now operational:**
  - **~200,000 patient records spanning 30 years**
  - **~120 million RDF triples**
  - **Used for research and quality reporting**
- **Average query speed improvement: ~264%**
  - **Measured over 1317 SPARQL queries**

Cleveland Clinic Semantic Strategies

- Build Centralized/federated semantic data repository
- **Define and collect stable core data elements and clinical facts**
- Define RDF data models augmented by domain and upper ontologies
- Link RDF instance data with ontologies and rules to support inference, query, and derived views

Experience: Core Data Elements

Why core data elements?

- **Data relativity - view of data dependent on frame of reference**
- **Temporal perspective: what is a pre-procedural risk factor from one point in time may be a post-procedural complication from another**
- **Definitional perspective: definitions for the same term can vary among uses and over time (e.g., current smoker)**
- **Version perspective: model/data versions**

Experience: Core Data Elements

How to define core data elements?

- **Event model: Most medical data can be easily organized into temporally discrete events with associated properties**
- **Fuzzy time: Timing of medical events can be fuzzy for many reasons. Need to embrace this fuzziness**
- **Pragmatic definitions: must find balance between infinitely reusable atomistic detail and special purpose definitions with limited reusability**

Experience: Core Data Elements

Strengths:

- **Multiple uses of the same data**
- **No need to collect and store the same data multiple times in different repositories for different purposes**

Experience: Core Data Elements

Challenges:

- **Poor alignment with current practice**
  - **Clinical practice is to document patient conditions anew with each encounter**
  - **Clinical documentation is part of legal record and cannot be changed once codified in patient medical record**
- **Past patient history**
  - **Data usually collected by clinicians from the perspective of the current encounter -- often lacks sufficient precision to convert to core data elements**

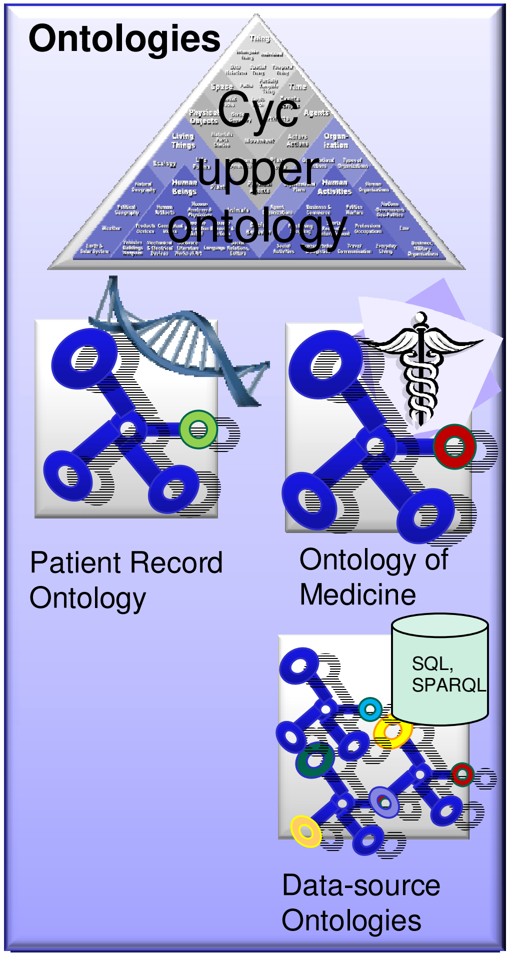
Cleveland Clinic Semantic Strategies

- Build Centralized/federated semantic data repository
- Define and collect stable core data elements and clinical facts
- **Define RDF data models augmented by domain and upper ontologies**
- Link RDF instance data with ontologies and rules to support inference, query, and derived views

Experience: Data Models & Ontologies

Semantic layers:

- **Patient record OWL ontology**
  - Used to express patient instance data
- **Bridging ontologies**
  - Align domain ontologies with term standards found in upper-level ontologies
  - Map from data source ontologies
- **Ontology of medicine**
  - Reference ontology of medical terms and relationships
- **Upper Ontology: Cyc**
  - General knowledge organization



Experience: Data Models & Ontologies

Strengths

- Provides a stable layer of terms through which to access instance data
- Supports different views of the same data

Challenges

- Lack of strong upper-level ontologies in medicine
- Maintenance of internal and external ontology alignments in the face of model changes

Cleveland Clinic Semantic Strategies

- Build Centralized/federated semantic data repository
- Define and collect stable core data elements and clinical facts
- Define RDF data models augmented by domain and upper ontologies
- **Link RDF instance data with ontologies and rules to support inference, query, and derived views**

Experience: Inference, Query and Views

Using inference to enhance queries & views

- **Forward inference:**
  - Inference run before query time
  - Either for persistence or on-the-fly use
- **Backward inference:**
  - Inference run at query time

Used to facilitate query formulation, data exporting, and report generation

Experience: Inference, Query and Views

Strengths:

- **Queries can be asked using terms not present in the instance data (using inference)**
- **Caching and periodic refreshing of different views of the data (e.g., an STS view, a SNOMED view, etc.)**
- **Allows maintaining different versions of the same view**

Experience: Inference, Query and Views

Challenges:

- **Inference performance bottlenecks:**
  - Not yet using Oracle's built-in inference engine
  - Currently using external inference engines
    - Combination of custom RETE inferencer, OWL, N3 rules Cyc inference engine and Python RDFlib
  - Forward inference is slow and degrades significantly as the number of graphs and the number of events per graph increase
- **Maintaining semantic alignment:**
  - Different versions of instance data, rules and ontologies must be kept in alignment as changes occur

Questions?

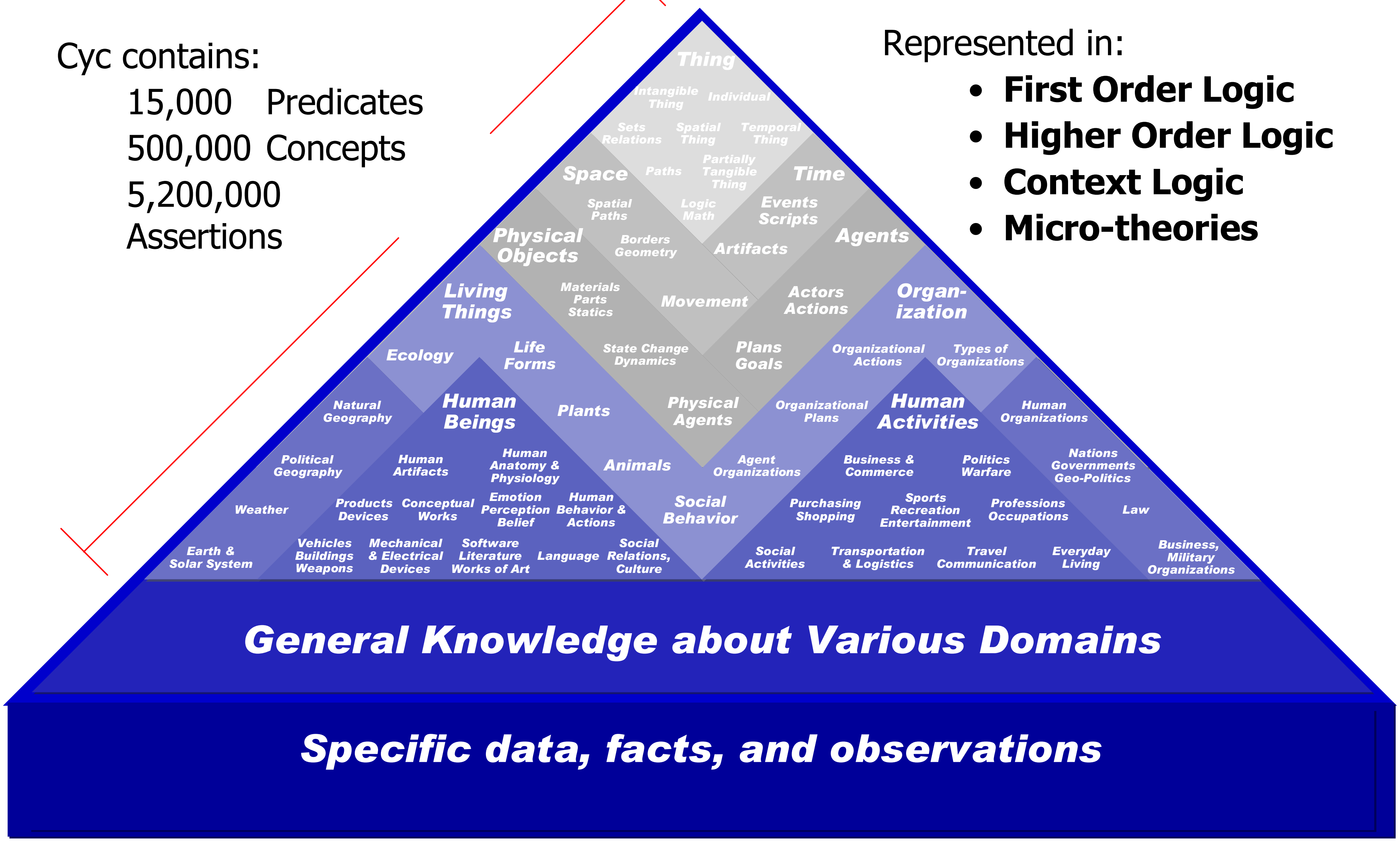
The Miller Family Pavilion houses the largest and most advanced cardiovascular care facility in the world.
Cleveland Clinic
Sydell & Arnold Miller Family Pavilion
"We are honored to care for our patients and honored to work with our colleagues. The goal for all of us working together is to achieve what is best for the individual patient."
— Bruce W. Lytle, MD
Chairman, Miller Family Heart & Vascular Institute
Sydell and Arnold Miller Family Heart & Vascular Institute
7

BACKUP SLIDES

Cyc Knowledge Base

Cyc contains:

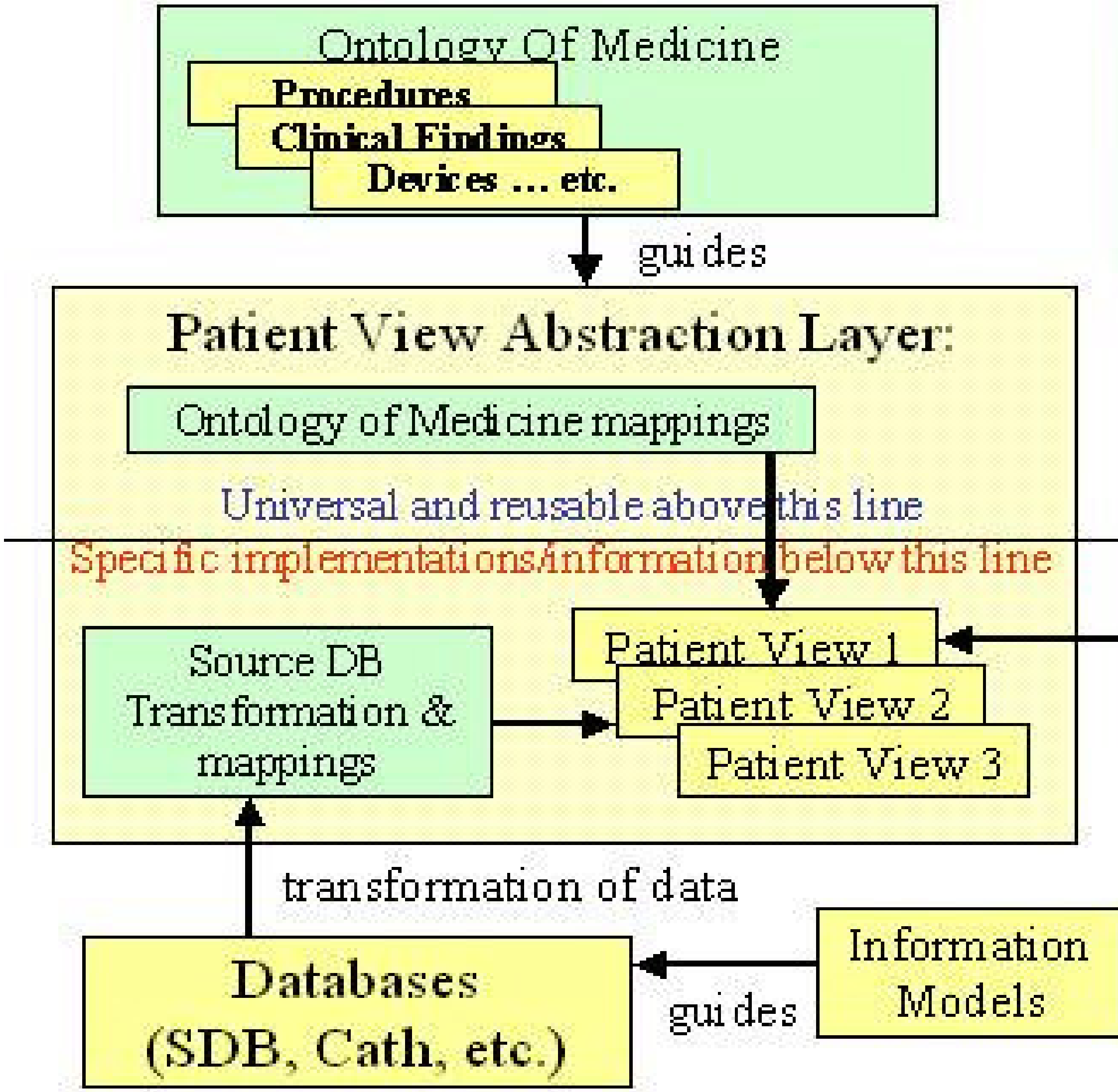
15,000 Predicates
500,000 Concepts
5,200,000 Assertions



Represented in:

- **First Order Logic**
- **Higher Order Logic**
- **Context Logic**
- **Micro-theories**

Experience: Data Models & Ontologies

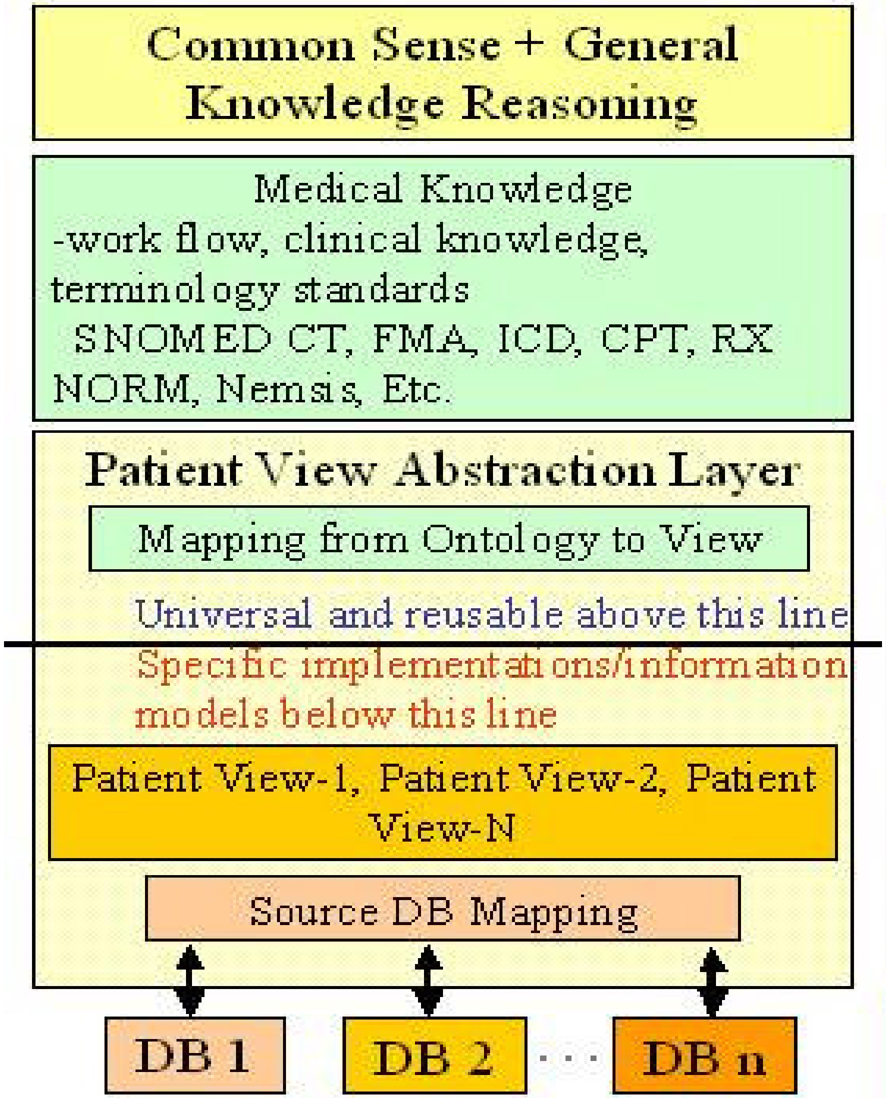


Ontology of Medicine: medical knowledge, mappings to standards: SNOMED, ICD, CPT, etc.

transformation of data in source DBs results in "patient views" used in research, external and internal reporting, data export, etc.

XML Schemata, XML pick lists Database models, Access mappings

Experience: Data Models & Ontologies



CYC: Upper and middle ontology supports: temporal reasoning, natural language query, report generation, etc.

Ontology of Medicine (OOM), Core Data Elements: universal, reusable: guides transformation, enables integration from disparate systems, enables term reuse, etc.

The Patient View Abstraction Layer defines transformations of source data. An implementation of this layer results in a dataset (view) that is aligned with the Ontology of Medicine and terminology standards. It consists of mappings from the OOM to views of the data and mappings from source data to views.

Source databases: CATH, ECHO, workflow systems, etc.