DSE 203

DAY 1: DATA INTEGRATION USE CASES

A Business Case (from IBM)

Suncorp is a diversified financial services group that offers general insurance, banking, life insurance and wealth management services. With operations in Australia and New Zealand, Suncorp has over AU$95 billion in assets, more than 16,000 employees and relationships with over nine million customers. The financial services organization maintains five operating divisions, managing 14 market brands, and is supported by corporate and shared services divisions.

"Mergers and acquisitions in the past decade have increased our customer base by 200 percent. Having a single view of the customer, we're more accurately able to target and cross-sell across our brands."

Suncorp-Metway Ltd wanted a single, integrated view of its customers to ensure its marketing campaigns didn't encourage internal conflict between the brands and duplication of efforts, both of which had a negative effect on the bottom line.

Deconstructing the Business Case – Hypothetically

Insurance Company's Partial Schema

Policies(PolicyKey, PolicyTypeKey, Agent, Conditions)
PolicySales(PolicyKey, PolicyholderKey, StartDate, TransactKey,Premium,CoveragePeriod, CoverageLimit)
Transactions(TransactKey, Date, Time, Amount, Balance)
Policyholders(PolicyHolderKey, Name, Address, City, State, ZIP)
Claims(PolicyKey, ClaimKey, TransactKey, ClaimAmount)
ClaimDescription(ClaimKey, TypeKey, ClaimantKey, ProcCode, Description)
Claimants(ClaimantKey, Name, Address, City, State, ZIP)
ClaimTypes(TypeKey, Description)
PolicyTypes(PolicyTypeKey, Name, Description)

Bank's Partial Schema

Accounts(AcctNumber, AcctType, MemberID, MemberType, TypeID, StartDate, EndDate, InterestRate, CreditLimit)
Individuals(MemberID, FName, MI, LName, SSN, Nationality, DoB, LegalStatus, FullAddress, Phone, PhoneType, Email)
Corporations(MemberID, Name, RegisteredAddress, CorporationType, Signatory1, Signatory2, DNBNumber, Phone, Email)
Transactions(TrID, AcctNum, Date, Time, TransactionType, Description, TransactionAmount, Debit/Credit, Balance, Payoff)
AccountType(TypeID, Name, Description)
TransactionTypes(Ttype, Name, Description)
Disputes(AccntNumber, DisputeID, TrID, Date, DisputeAmt, Explanation, Valid, ValidatorID)

Deconstructing the Business Case – Hypothetically

- What would the company like to do when these databases are integrated?
  - Find customers who should be given a discount for their insurance because they also bank with the company
  - Which "high-risk" customers have we insured, based on
    - Duration of their association with us, claims they have filed, credit and checking account history, frequency of disputed transactions
  - Which customers who had auto insurance with us in the past but do not insure through us now, have current auto-loans through us? Do a time-based analysis to find out if there were sharp dropping points ?
- Queries and Analytics – two somewhat different needs of enterprise integration systems

Integration of Public Health Information

Washington D.C. Department of Health Automated Disease Surveillance System

After the postal anthrax crisis in 2002, DC DOH was under intense scrutiny and the agency felt responsible to serve as a model for the rest of the country. The DC DOH realized it had systematic problems, including information silos, long delays in obtaining clinical information, and incomplete planning for known and unknown hazards.[4] All case reports were submitted manually, on paper, via mail or fax, which was time-consuming and error-prone for both the submitters and the DC DOH epidemiologists receiving the information. If hospital staff were not working during holidays, weekends, or evenings, information was inconsistently captured and reports could be missed. As a result, DC DOH could not be confident it had complete information and that its investigations were accurate and prompt.

- Create an effective bioterrorism and disease surveillance system through integrated public health infrastructure
- Integrate
  - Hospitals and Labs
  - Poison control
  - Emergency services
  - Animal disease control
- Provide
  - Continuous public health data collection
  - Real-time monitoring and analysis
  - Early monitoring and intervention

Integration of Public Health Infrastructure

Washington DC Disease Surveillance System (WADDS)

CATEGORY	SOLUTION	DESCRIPTION
Data Exchange	Integration hub with HL7 messaging	All internal and external data moves through commercial integration hub that transforms HL7 V2 data into a consistent HL7 V3 representation.
Terminology	SNOMED LOINC	Implemented standard concept terminologies SNOMED and LOINC for coding of clinical and lab data.
Conceptual	RIM-based integrated data repository	A centralized, commercial data repository was natively designed on the HL7 RIM to normalize clinical data from disparate sources. Implemented a data quality algorithm to manage patient matching and identify duplicate records.
Architecture	PHIN architecture	Developed architecture consistent with the CDC's Public Health Information Network requirements.

*Used to enable interoperability between existing hospital and lab systems and WADSS.

Reference Information Model

Data exchange is the process of taking **data** structured under a source schema and transforming it into **data** structured under a target schema, so that the target **data** is an accurate representation of the source **data**.

What does the integration involve? HL7

Health Level-7 or **HL7** refers to a set of international standards for transfer of clinical and administrative data between software applications used by various healthcare providers.

- HL7 comes in two forms – a provider can choose either

```
MSH|^~\&|LAB|767543|ADT|767543|19900314130405||ADT^A04|XX3657|P|2.3.1<CR>
EVN|A01|19980327101314|19980327095000|I||19980327095000<CR>
PID|1||123456789ABCDEF|123456789ABCDEF|PATIENT^BOB^S||19590520|M||
    612345 MAIN STREET^^ANYTOWN^CA^91234||714-555-1212|
    714-555-1212||||123456789ABCDEF|||U<CR>
PD1||||WELBY<CR>
PV1|1|0||NEW|||SPOCK<CR>
```

```
<!DOCTYPE ADT_A03 SYSTEM "hl7_v231.dtd">
<ADT_A03>
<MSH>
        <MSH.1>|</MSH.1>
        <MSH.2>^~\&amp;</MSH.2>
        <MSH.3><HD.1>LAB</HD.1></MSH.3>
        <MSH.4><HD.1>767543</HD.1></MSH.4>
        <MSH.5><HD.1>ADT</HD.1></MSH.5>
        <MSH.6><HD.1>767543</HD.1></MSH.6>
        <MSH.7>19900314130405</MSH.7>
        <MSH.9>
                <CM_MSG_TYPE.1>ADT</CM_MSG_TYPE.1>
                <CM_MSG_TYPE.2>A04</CM_MSG_TYPE.2>
        </MSH.9>
        <MSH.10>XX3657</MSH.10>
        <MSH.11><PT.1>P</PT.1></MSH.11>
        <MSH.12><VID.1>2.3.1</VID.1></MSH.12>
</MSH>
<EVN>
        <EVN.1>A01</EVN.1>
        <EVN.2>19980327101314</EVN.2>
        <EVN.3>19980327095000</EVN.3>
        <EVN.4>I</EVN.4>
        <EVN.6>19980327095000</EVN.6>
</EVN>
<PID>
        <PID.1>1</PID.1>
        <PID.3.LST>
                <PID.3><CX.1>123456789ABCDEF</CX.1></PID.3>
```

- Queries
  - Find all prescriptions and lab reports of patient #19590520 containing serum protein, along with age-specific normal values between 1/1/2012 and 9/1/2015
    - The patient went to three different clinics and four different labs in this period
    - The doctor's own office uses a relational database for EHR

What does the integration involve? LOINC

- **Logical Observation Identifiers Names and Codes (LOINC)** is a database and universal standard for identifying medical laboratory observations.

2. COMPONENT	Text	255	First major axis-component or analyte
3. PROPERTY	Text	30	Second major axis-property observed (e.g., mass vs. substance)
4. TIME_ASPCT	Text	15	Third major axis-timing of the measurement (e.g., point in time vs 24 hours)
5. SYSTEM	Text	100	Fourth major axis-type of specimen or system (e.g., serum vs urine)
6. SCALE_TYP	Text	30	Fifth major axis-scale of measurement (e.g., qualitative vs. quantitative)
7. METHOD_TYP	Text	50	Sixth major axis-method of measurement
8. CLASS	Text	20	An arbitrary classification of the terms for grouping related observations together. The current classifications are listed in Table 32. We present the database sorted by the class field within class type (see field 23). Users of the database should feel free to re-sort the database in any way they find useful, and/or to add their own classifying fields to the database. The content of the laboratory test subclasses should be obvious from the subclass name.

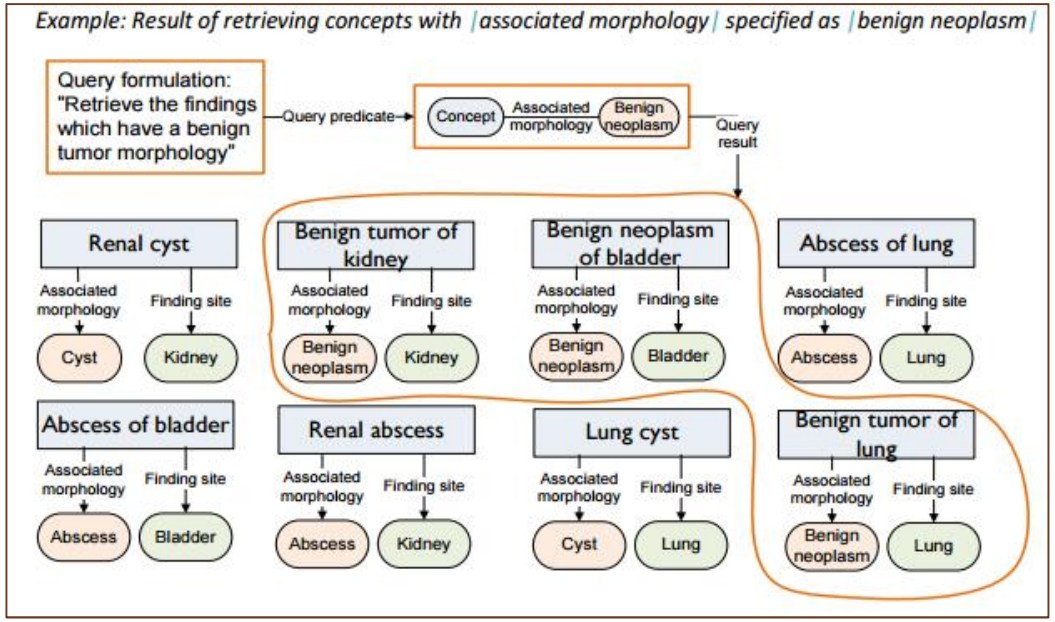
BP.ATOM	Blood pressure atomic
BP.CENT.MOLEC	Blood pressure central molecular
BP.MOLEC	Blood pressure molecular
BP.PSTN.MOLEC	Blood pressure positional molecular
BP.TIMED.MOLEC	Blood pressure timed molecular
BP.VENOUS.MOLEC	Blood pressure venous molecular
CARD.RISK	Cardiac Risk Scales Framingham
CARD.US	Cardiac ultrasound (was US.ECHO)
CARDIO-PULM	Cardiopulmonary
CLIN	Clinical NEC (not elsewhere classified)

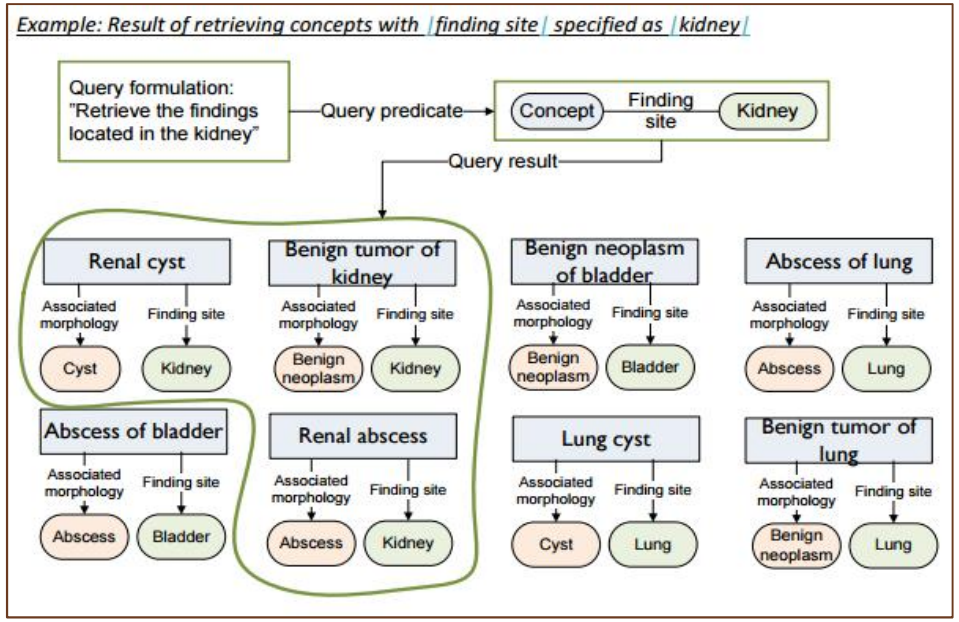
MOLPATH.DELDUP	Gene deletions or duplications
MOLPATH.INV	Gene inversion
MOLPATH.MISC	Gene miscellaneous
MOLPATH.MUT	Gene mutation
MOLPATH.REARRANGE	Gene rearrangement
MOLPATH.TRINUC	Gene trinucleotide repeats
MOLPATH.TRISOMY	Gene chromosome trisomy
MOLPATH.TRNLOC	Gene translocation

Some Attribute Domains are hierarchical

What does the integration involve? SNOMED

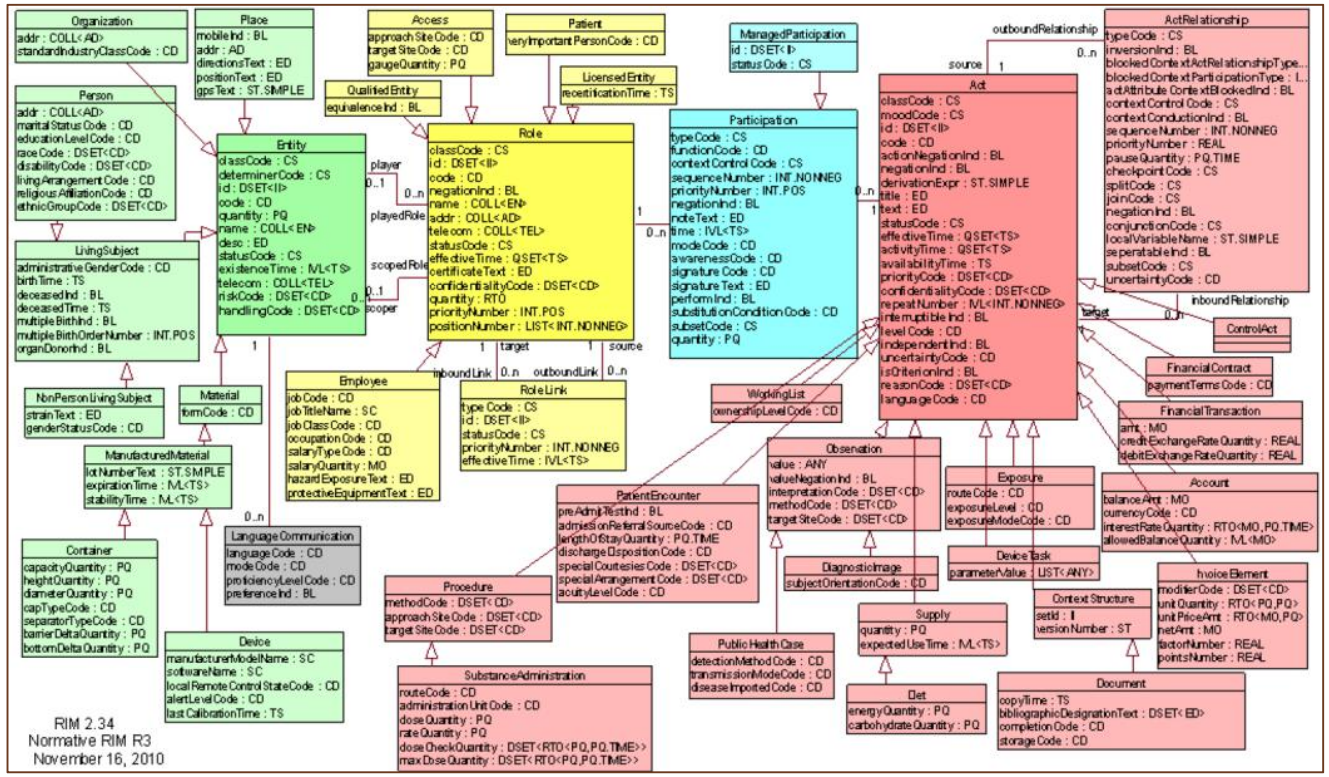
- The Systematized Nomenclature of Medicine (**SNOMED**) is a systematic, computer-processable collection of medical terms, in human and veterinary medicine, to provide codes, terms, synonyms and definitions which cover anatomy, diseases, findings, procedures, microorganisms, substances, etc.





Issues in the Healthcare Use Case

- Integration across multiple models of data
- Global Schema – RIM
- Format conversions
- Data Exchange
- Terminology Integration
  - What happens if there are conflicts?



Integration for Multichannel Customer Analytics

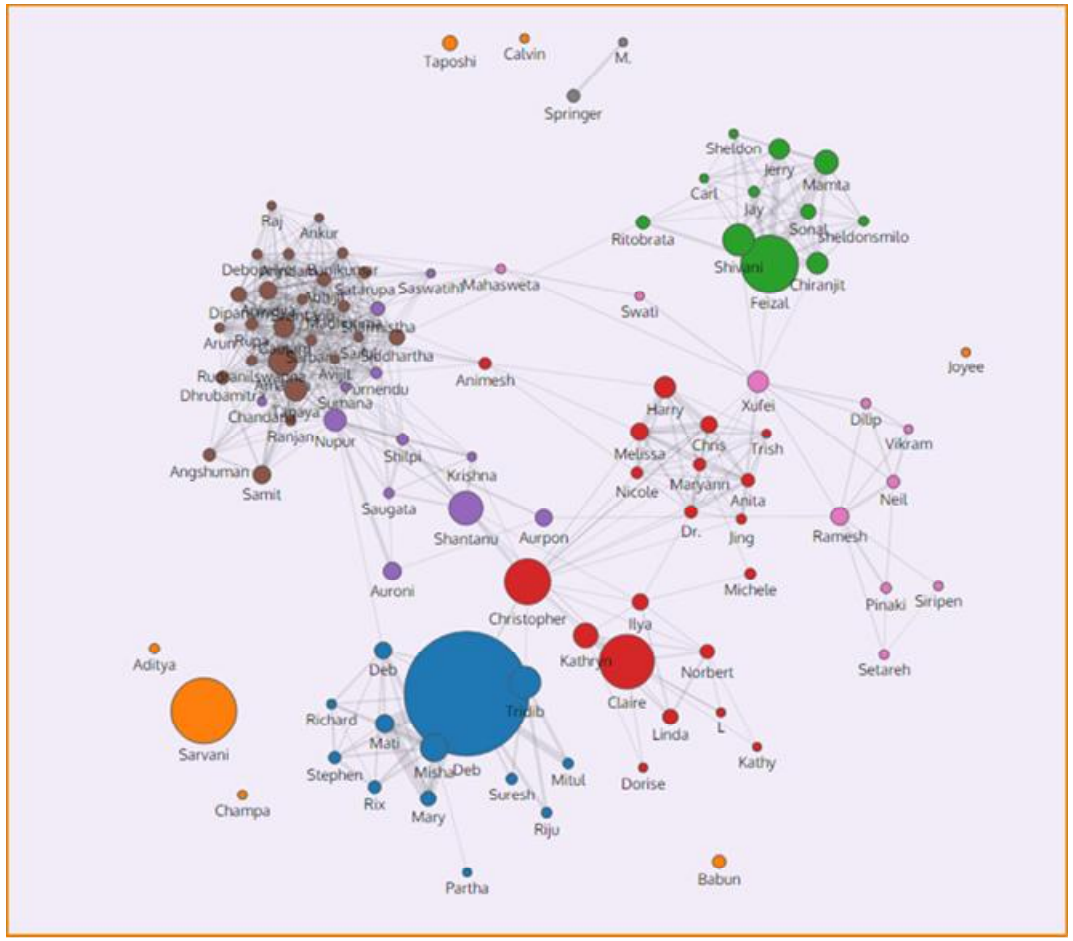
- Customer analytics
  - processes and technologies that give organizations the customer insight necessary to deliver offers that are anticipated, relevant and timely
- Questions one would like to ask
  - Is our product launch going well?
  - Is there an emerging product issue?
  - Where should the product team focus its development dollars?
  - Are there more effective methods for positioning current products?
  - Which services have the best chance of surviving a turbulent market?
  - Is there a product defect in the market?

What does the integration involve? Document & Text Data

- Document data is usually semi-structured
  - Text – unstructured part of document data
- Text is can be accessed indirectly for integration
  - The integration system accesses the output of a text analysis software
    - Sentiments, named entities, co-references, relationships
    - This output can be treated as structured data
- Text can be "structured"
  - Based on knowledge of the document and "domain knowledge"
    - Variable accuracy and success
- Integration Query
  - What topics are being recently discussed about product X?
  - What other products are being discussed more favorably than ours?

What does the integration involve? Graph Data

- Captures entity instances and their relationships to others
  - Email/Social Networks
    - Who receives how many messages from whom
    - Threads and how they evolve
    - Time-varying connections
    - Events
  - Facebook Graph API
- Integrated query
  - Within the k-level followers of Mr. X, who are giving bad publicity to our audio products that use the amplifier model Y?



What does the integration involve? Semantics

- Entity Matching
  - Users use different identifiers in different systems
  - Record Linking
    - Can we match the identities of users by comparing the rest of their data records?
- Semantic Constraints
  - Conditions (predicates) that hold for a data schema
  - How do the semantics of the data imposed at the source relate to the semantics of the integrated data?
  - Ontological constraints
    - Constraints that are not explicitly specified at the sources but hold in the domain
      - Can they be used for integration and query processing?

Tailpiece

- There is a wide variety of data integration scenarios and related issues
- Many modern information integration systems need to perform the integration across data models
- We will spend most of our time on relational-like systems
- Things outside our scope
  - Manual Data Integration
  - Common User Interface
  - Integration by Applications
  - Common Data Storage
  - Data Integration using Workflow Management