Databases in Radiotherapy and a brief introduction to SQL

• Peter Balter, Ph.D.

What is a database

• It is an organized collection of data
  – Could be a paper file system
  – Could be spread sheet
  (but neither are very good databases)
Databases in Radiotherapy

- Many of the systems in Radiotherapy are primarily databases with specialized equipment and user interfaces
  - MOSAIQ
  - Aria
  - Sun Nuclear Atlas
  - Raystation
- Other systems are not primarily database but generate a large amount of data that is organized using a database
  - Varian RPM
  - Radcalc
  - Pinnacle

Relational Database

- Data is grouped into tables based on its content
- The groupings have known relationships to each other
- Data access is done using these relationships
**Relational Database**

- Databases contain one or more tables
- Tables contain one or more fields

**Post Relational Databases (noSQL)**

- Considered by many to be a superior way to manage data
- Each record (document) carries its own fields and values with it (keys)
- Requires much less up-front work than a relational database
  - Does not require that relationships be defined between data elements
  - Allows modifications of database on-the-fly (no pre-defined schema)
- Showing up in many new systems
  - MOBIUS use MONGO and open source noSQL
Relational Database: Tables

- Each table generally relates to a real-world entity
  - Example: all contact information for an institution may be grouped in a single table

Relational Database: Fields

- Each record is stored in a table as a collection of fixed length fields
- Each field has a data type and a length
  - Similar to data type in programming languages
  - float, real, char, text, date
  - Each takes up a fixed amount of storage space
- In the example on the right ZIP code is stored as Text with 50 characters rather than numeric with 7 digits to allow free-form entry of foreign zip-codes (also worked well when +4 zips codes were introduced)
- Data validation can occur by database design or by rules enforced on entry
Primary Keys

- Most tables have a primary key (field) that is used to define the relationship between that table with other tables
- The primary key can be the composite of 2 fields
- Often the primary key is a sequence number rather than a physical characteristic of the entity
  - Example: The primary key on a person could be their name but names change or could have been miss-spelled on entry so most databases may have a patient number usually independent of the hospital patient ID

Example from Varian Unified Reports Application Schema

---

Foreign Keys

- Foreign Keys are fields in a table that point to the Primary Key of another table
  - Are restricted to being available in the other table to enforce data integrity

In this example the primary key for PatientInVivoDatModel is also a foreign key since inVivo dosimetry must be related to a unique patient

Example from Varian Unified Reports Application Schema
Database Schema/Data Dictionary

• Database Schema
  – The blueprint for the database
  – Shows how data is divided into tables and how tables relate to one another
  – Shows integrity constrains
  – Shows stored procedures

• Data Dictionary
  – Should include much of the same information as the schema
  – May contain further information
    • Detailed descriptions of the data

Typical Database Schema (in-house Pinnacle add-on database)
Data Dictionary(ARIA) - Public

Data Dictionary(MOSAIQ) - Available but not Public
Data Dictionary – other systems

- Most systems do not have publicly available data-dictionaries
  - Few users are requesting them
- Most systems can be reasonably reverse engineered
  - Browse the data and table with an Database tool
  - Try and match data in tables to that in the GUI
- Many vendors will help with queries
  - The examples from Sun Nuclear Atlas were mostly hacked
  - Understanding how to re-derive data not stored directly in the database required help from the Vendor (which they provided)

Relationships:

- One-to-Many
  - A row in one table can have a large number of matching rows in another table
  - This is the most common relationship
  - Example: A single plan can have many beams but each beam can only belong to a single plan
- Many-to-Many
- One-to-One
Relationships:

- One-to-Many
- Many-to-Many
  - A row in one table can have a large number of matching rows in another table and vise-versa
  - Example: Plans can have multiple imagesets and imagesets can belong to multiple plans
- One-to-One

Relationships:

- One-to-Many
- Many-to-Many
- One-to-One
  - Each row in one table can have exactly one matching row in another table
  - Not common as if this relationship exists the rows could have been placed in the same table.
Queries and SQL

• SQL (SEQUEL): Structured English Query Language
• The standard means of retrieving data from a relational database
• Is very powerful but can be used simply
• The standard form is
  – Select <fields> from <table> where <criteria>;
• Queries can returned sorted results
  – Select <fields> from <table> where <criteria> ORDER BY <field>;

Simple Query Example MOSAIQ

Find patients internal ID based on their Hospital Medical Record Number

• Select ident.Pat_Id1, ident.IDA from ident where ident.IDA='ZZZCIRS'

Implemented in MicrosoftSQL Management Studio (but should work in any query engine that supports SQL)
Simple Query Example: Pinnacle

- Uses the built in PostgreSQL interactive terminal to find the number of patients in the active database
  
  ```
  \psql -d p3rtp -h localhost -U lpuser
  Welcome to psql 8.3.5, the PostgreSQL interactive terminal.
  ```
  
  - Type: `\copyright` for distribution terms
  - `\h` for help with SQL commands
  - `\?` for help with psql commands
  - `\g` or terminate with semicolon to execute query
  - `\q` to quit

  ```
  p3rtp=> SELECT COUNT("patientid") FROM "patient";
  count
  ------
  1193
  (1 row)
  ```

Micheal Kantor
Example Query (Sun Nuclear Atlas)

- SELECT `dbo_Machine`.MachineName FROM `dbo_Machine` returns an unsorted list of machines in the database
- SELECT `dbo_Machine`.MachineName FROM `dbo_Machine` ORDER BY `dbo_Machine`.MachineName; returns a sorted list

Queries and SQL using Joins

- Queries can span multiple tables using Joins
  - Select `<fields>` from `<table>` Join `<table2>` on `<common field>` where `<criteria>`;
  - These tables can span different databases
- Joins can be Inner or Outer Joins
  - Inner joins (default join) only returns data that has matching rows in both tables
  - Outer joins will give data that exists in either table
- Example if you have a table of equipment and one of calibrations and inner join on equipment will only give equipment with calibrations an out join will give all equipment if if no calibration exists
Simple Query Example MOSAIQ with Join

Find patients internal ID based on their Hospital Medical Record Number and get their name as well (stored in another table)

Select ident.Pat_Id1, ident.IDA, Patient.Last_Name, Patient.First_Name from ident
inner join patient on ident.Pat_Id1=patient.Pat_ID1
where ident.IDA='ZZZCIRS'

Implemented in Microsoft SQL Management Studio (but should work in any query engine that supports SQL)
Example Join Query (Sun Nuclear Atlas)

- If we want to know what energies each machine has we need to use joins since that information spans 3 tables

Example done using Microsoft Access to Query SQL Database

Queries and SQL using aggregate functions

- Queries can return summary data (SQL aggregation functions)
  - Count, Sum, Max, Min, Avg, etc
  - Example: Select avg(<field A>) from <table> where <criteria>
    Gives the average value of <field A>
- Generally used with the “GROUP BY” clause to define what set of data is being aggregated
  - Example: Select <field B>, avg(<field A>) from <table> where <criteria> group by <field B>
    Gives the average value of <field A> for each <field B>
Example aggregate Query (Sun Nuclear Atlas)

- We can use aggregate functions if we want to know how many of each energy we have across all machines:

Example aggregate query using a derived value (Sun Nuclear Atlas)

- Since Energy = 6 can be 6 MV or 6 Mev to further specify we need to build a derived field to sort on that includes both the energy value and the units
- We can query and sort based on this derived value: EnergyValue & [Units]
4/6/2018

Query Building tools

• Building more useful queries like “what is the average output and symmetry for each machine by energy” can be very complicated

  Avg(dbo_Dqa3Data.ABDose) AS AvgOfABDose, Avg(dbo_Dqa3Data.ABAxialsym) AS AvgOfABAxialsym,
  Avg(dbo_Dqa3Data.ABTrsym) AS AvgOfABTrsym FROM (((dbo_Machine INNER JOIN dbo_MachineEnergy ON
  BY dbo_Machine.MachineName, [dbo_Energy][EnergyValue] & [dbo_Energy][Units];

Query Building tools

• There are a large number of query building tools that have GUIs that can help you quickly build complicated queries
Queries that change the data

- Some query keywords can be used to change the data in the database
  - Insert: adds rows
  - Update: modifies rows
  - Delete: removes rows
  - Merge: combines rows
- These can be very powerful but:
  - Can unexpectedly modify/remove large amounts of data
  - Should not be used on clinical databases (ARIA, MOSAIQ)

Example Aria: Scripting API for data access

- Many programming languages have built-in support for data access
- The Varian API provides objects\methods\structures to query the ARIA/Eclipse database
  - // getting all treatment plan’s data under “C2” course in selected patient
  - var allPlansInfo = from Course c in ThePatient.Courses
    where c.Id == "C2"
    select new
    { plans = c.PlanSetups,
      Course = c, };
  
  //query finds the first PTV structure:
  
  Structure target = (from s in StructureSet.Structures
    where s.DicomType == "PTV"
    select s).FirstOrDefault();

Amy Liu
Queries that return a large amount of data

• Poorly written (or thought out) queries can adversely affect the performance of the database server the one that is also running your clinic in the case of MOSAIQ or ARIA)
  – Queries should have a “LIMIT” clause during testing if supported by your SQL server
  – Queries should be tested on non-clinical systems first, if possible.
    • If no development system exists consider testing during off-hours
  – For relatively small databases consider making a backup of the database and run queries against the backup during debugging

GUI based Query builders

• There are multiple sets off software both commercial and open source that will provide a framework for helping you build queries.
  – Many will automatically provide joins based on the database scheme
  – The will write SQL than can be manually edited, if needed, and/or copied into code for automated querying

• Examples
  – Microsoft SQL management studio
  – MS Access (used for some of the examples in this work)
  – pgAdmin (PostgreSQL)
  – Many others (check google)
Views, Triggers, and Stored procedures

- Most SQL servers support the serval nondate objects in the database used for efficient data access and database integrity:
  - Views: Predefined queries to simplify the users access to the database
  - Stored Procedures: Pre-compiled SQL statements
    - Run faster than normal queries
  - Triggers: Procedures designed to run automatically based on other events
    - Can be used for data integrity checks
    - Often used to create audit logs

Why bother writing you own queries

- Most or all data achievable in the query is available via the vendors GUIs
- Consolidating data from the GUI can be laborious
- Trend analysis, trivial in the database space, can be impractical via GUI
  - Example: How many times in the last year was a film rejected by MD in the last year

Select image.Att_App, count(image.Att_App) from image where (image.Image_Class=9 or image.Image_Class=13 ) and image.Study DtTm > '2017-01-01' group by image.Att_App

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Reviewed</td>
<td>4,440</td>
<td>1.2%</td>
</tr>
<tr>
<td>Approved</td>
<td>324,541</td>
<td>88.4%</td>
</tr>
<tr>
<td>Rejected</td>
<td>38,257</td>
<td>10.4%</td>
</tr>
<tr>
<td>Total</td>
<td>367,238</td>
<td></td>
</tr>
</tbody>
</table>
Reports

- Forms with areas filled in by queries embedded within the form
- Often the only “official” way to get summary data out of many systems
- Are useful when the same data is to be queried and presented multiple times
  - Physics weekly check list from MOSAIQ
  - Billing documents
  - Treatment summaries
- Some database systems have integrated report generators
  - MS Access
  - Microsoft SQL Server Reporting Services (SRSS)
- There are many 3rd party report generators
  - Crystal Reports is a combination query builder and report generated used as the OEM report generator for many systems

The Client Server Database Model

- The database runs as it’s own process either on the local computer or on a distant computer
- Applications send transactions to the database
- Transactions maybe processed via an intermediate layer (driver)
  - Microsoft ODBC driver allows many different types of databases to talk to one another

![Database Diagram]

**Client** — TCP/IP — **Server Process** — **Database**
Advantages of the client server model

- Separates database development from application development
- Client applications are independent from physical location of data
- Client systems can be optimized for display and user interface while database server can be optimized for performance
- Reduce traffic on the network as only processed data is transferred
- Can handle concurrent access by multiple users (better than a shared file)

Common Database systems

- Oracle: High performance “main frame” database
- Microsoft SQL: High performance clustered system or local “lite versions” available
- PostgreSQL: Open source database
- MySQL: Another very popular open source database
  - RedCap (front end on MySQL) for “building and managing online surveys and databases”
- MS Access: great general purpose database that hides much of coding.
- SQLite: open source database used as an embedded database in many other applications
- Many many more (just google open source SQL)
### Databases in Radiotherapy

- MOSAIQ: Microsoft SQL
- Aria
- Sun Nuclear Atlas: Oracle
- Raystation
- FUJI Evercore: SQLite
- Radcalc
- Pinnacle (for a limited amount of the data): PostgreSQL
- Varian RPM: mdb file (MS Access)

Note the SQL formalism is general enough that many system can support more than one type of database.

### Backup

- Database systems have backup utilities that the end user should be able to use
  - Simple systems may be just a file backup
  - Other systems have backup and restore functions within the SQL server workspace
  - Many systems exist for real-time (or near-time) backup between database clusters
  - Many enterprise systems allow “rewind” back to a state at an earlier time
- Work with your vendor to understand what backup systems they support and how to best implement backups in your environment
Audit logs

- Many SQL systems have built in auditing
  - Can tell when the database schema has changed
  - Can tell when data has been
    - Accessed
    - Changed
    - Added
    - Deleted
- Many databases have built in history logs that can be used for auditing
  - MOSIAQ keeps previous versions of many table entries in the database
    - Can produce unexpected query results
  - Some are not easily accessible from the GUI but can be found when directly querying the database – Has helped with some root-cause analysis investigations.

- Great thanks to Michael Kantor for many of the examples in this talk as well as a great expansion of the content
- Thank you to Amy Liu for Eclipse API examples