DATA MODELS FOR A
REGISTRAR’S DATA MART

by

RONALD GAGE ALLAN
Office of Student Financial Services
Georgetown University
Washington, DC
allanr@gunet.georgetown.edu

and

DAVID R. MAY
Vice President
Xpedior Inc.
David.May@xpedior.com

ABSTRACT: The purpose of this paper is to address the issues associated with the modeling of student record data for use in the Student Record Data Mart portion of the institution’s Data Warehouse. Generalized data models of Registrar related data in the Dimensional Modeling format are provided. These models were developed at Georgetown University which operates under a semester system, keeps student records in a SIS-Plus computer system, follows a Need-Blind Admissions policy and a Full-Need Financial Aid policy. Every effort has been made to generalize the models but users at other institutions will have to adapt them to their local circumstances.

Data warehousing has come into being because the file structure of the large mainframe core business systems is inimical to information retrieval. The purpose of the data warehouse is to combine core business and data from other sources in a format that facilitates reporting and decision support. In just a few years, data warehouses have evolved from large, centralized data repositories to subject specific, but independent, data marts and now to dependant marts that load data from a central repository of Data Staging files, that has previously extracted data from the institution’s operational business systems (e.g., student record, finance and human resource systems, etc.). The Dimensional Model is a logical design technique that seeks to make data available to the user in an intuitive framework that is intended to facilitate querying (Kimball, 1997).
DATA MODELS FOR A REGISTRAR’S DATA MART

The purpose of this paper is to address the issues associated with the modeling of student record data for information access purposes. Generalized data models are provided. These models were derived at Georgetown University which operates under a semester system, keeps student records in a SIS-Plus computer system, follows a Need-Blind Admissions policy and a Full-Need Financial Aid policy. Every effort has been made to generalize the models but users at other institutions will have to adapt them to their local circumstances.

BACKGROUND

To date, little work has been done on fitting student record data to the dimensional model star schema. This is probably for the same reason that student record systems tend to be so poorly developed compared to accounting and human resource systems: the potential client base for student record systems in the United States is 3,700 colleges and universities versus a potential client base for accounting systems of 4.6 million corporations. The smaller client base has meant that far less effort has gone into the development and standardization of student record systems than is the case for accounting systems. As a result, the management and maintenance of student record systems within the institutions that use them tends to be significantly more resource and attention intensive than accounting systems, and significantly more ad hoc.

REQUIREMENTS OF A STUDENT RECORD REPORTING SYSTEM

Before establishing the requirements of a Student Record, or any other, reporting system, it is well to review what a data warehouse is and why data warehousing is done.

What is a Data Warehouse

A data warehouse is a “subject oriented, integrated, nonvolatile and time variant collection of data in support of management’s decisions” (Inmon, 1996). In Higher Education it is usually placed on a server different from the server on which the daily processing systems are maintained in order to minimize the impact on daily production processing.

Why Data Warehousing is Done

Data warehousing has come into being because the file structure of the large mainframe core business systems is inimical to information retrieval. The purpose of the data warehouse is to combine core business and data from other sources in a format that facilitates reporting and decision support.

A subtle distinction, but one that should be borne in mind, is that a Student Record Data Mart (SRDM) has a much higher operational component than have more typical higher education reporting systems. Interest in developing a SRDM tends to be high because Student Record processing is so complex that the core business systems are unusually grudging in giving up useful information on other than a student by student basis.

Accordingly, the authors propose a basic set of reporting requirements, that emphasize operational processes, for readers to use as a point of departure for developing their own requirements. These requirements are referred to as the Pierce Requirements after John Q. Pierce, the Georgetown Registrar who first articulated them. It is the presumption of the authors that the emphasis on the operational nature of the SRDM will be offset by the construction of an Institutional Research Data Mart that will emphasize responses to the more “big picture” types of queries required by the institution’s upper management.

The Pierce Requirements

The Data Warehouse in general, and the Student Record Data Mart in particular, should:

1. Provide adequate reporting capability on three levels:
   --> Executive
   --> Managerial
   --> Operational
2. Provide the maximum amount of ad hoc query capability; i.e., the data warehouse should be designed to minimize
the amount of advance planning and research required before the user commences the query and should not be designed to favor one type or set of queries at the expense of another type or set of queries.

3. Be designed such that the user needs no knowledge beyond his or her business processes to execute a query successfully (i.e., no special knowledge of the file structure or the reporting tool).

Although the Pierce Requirements are meant to emphasize operational capabilities, the third requirement, referred to locally as the Pierce Requirement, stands for the proposition that the data mart should be made as intuitive to the users as possible. The potential of the Dimensional Model and the star schema to do so in the principal motive driving the development of this paper.

STUDENT RECORD SYSTEMS CONSTITUENCIES

Student record systems tend to have four basic components that serve four basic constituencies.

- Admissions
- Registrar
- Financial Aid
- Student Accounts

The activities and procedures of these four functions are very different; the only thing they really have in common is that their primary purpose is to provide services to students. Also, the activities of the other three functions tend to be “downstream” from those of the Registrar, making them dependant on Registrar data while the Registrar, with the exception of initial entry of student data into the system by the Admissions function, is fairly independent of the data of the other functions.

Admissions

The Admissions function consists of the recruitment of prospective students through market analysis, image and strategy development, communications, contact tracking and advising. It also includes receiving application information, monitoring and communicating applicant status, review, and yield management (the encouragement of admitted students to actually enroll.)

Registrar

The Registrar function includes developing and publishing the course schedule, registration of newly admitted and continuing students, tracking academic progress, conferring credentials, producing transcripts and providing information to facilitate advising.

Financial Aid

The Financial Aid function delivers financial assistance to students in the form of loans, employment, scholarships and grants. It informs and educates prospective and current students by counseling them and their parents as to how to pay for college or graduate school. It contributes to policy deliberations within the school to that ensure that the school remains affordable to the vast majority of its desired students; it maximizes the integration of federal, state and private Financial Aid funding as well as institutional funding. It has fiduciary responsibilities for administration of institutional, federal, state and private resources and must ensure that limited funds are equitably distributed in compliance with myriad federal and state regulations.

Student Accounts

The Student Accounts Office is responsible for receiving, recording and billing tuition, course fees, room and board, housing deposits, parking, fines, and payments on accounts. It also receives funds from the Financial Aid Office, applies those credits to the balance owed, and sends bills for the remainder due. Additionally, it processes loan excess checks and performs calculations associated with withdrawal.

DATA WAREHOUSE FUNDAMENTALS
In just a few years, data warehouses have evolved from large, centralized data repositories to subject specific, but independent, data marts and now to dependant marts that load data from a central repository of Data Staging files, that has previously extracted data from the institution’s operational business systems (e.g., student record, finance and human resource systems, etc.). The Staging Files typically collect data from all the business systems nightly, cleanse the data, and store the data in a normalized data format. The dependant data marts, in turn, receive data from the Staging Files, storing it in dimensional format for query and analysis purposes. See Diagram 1.

**THE DIMENSIONAL MODEL**

The dimensional model is a logical design technique that seeks to make data available to the user in an intuitive framework that is intended to facilitate querying (Kimball, 1997). Although the dimensional model is a relational model, it is not normalized and is subject to some important restrictions. Every dimensional model is composed of one table with a multipart key, called the fact table, and a set of smaller tables called dimension tables. Each dimension table has a single-part primary key that corresponds exactly to one of the components of the multipart key in the fact table. Because of this image of a central fact table surrounded by a group of dimension tables, the construct is sometimes called a star schema.

**FACT TABLES, DIMENSION TABLES AND GRAIN**

At this point the reader who is new to data warehousing is probably wondering what is the difference between a fact table and a dimension table? In general facts are those attributes, usually quantitative, that users wish to measure about a subject. Examples of facts in a student record context are numeric grades, number of courses taken and number of students enrolled.

Dimension tables consist of sets of highly correlated descriptive attributes that can be placed in an obvious category. For example, student record users often wish to measure the number of students enrolled by degree, college and major. In this case degree, college and major are dimensional attributes and can be placed in dimensions called Term, when referring to the current semester, or Matriculation, when referring to the overall course of study in which the student is enrolled.

The level of detail available in a star schema is referred to as its **grain**. Grain is the level of detail in all dimensions of a star schema, nominally the same as the level of detail in the star schema’s fact table. The level of detail in each dimension table is also referred to as its grain. For example, the granularity of a star schema for commercial purposes, with dimensions of time (year, quarter, month, day), store (region, district, store) and product (manufacturer, brand, product), would be “product sold in store by day.” The grain of a student record star schema with dimensions of Time (Academic Year, Term), Student Bio/Demo (one record per student) Term (one record per student per Term) and Student Matriculation (one record per student per Course of Study undertaken) would be “Student per Term per Course of Study.”

**PRIMARY KEYS, FOREIGN KEYS, CONCATENATED KEYS, SURROGATE KEYS AND REFERENTIAL INTEGRITY**

A **primary key** is a unique key for each record in a dimensional table. The primary key of each dimension in a star schema is replicated in its fact table where it is referred to as a **foreign key**. When each primary key is repeated as a foreign key in the appropriate record in the fact table, the star schema is said to have **referential integrity**. Because Registrar related keys are denominated by Terms and Financial Aid keys by Award Years, referential integrity makes it difficult for the two constituencies to share dimensional tables.

The foreign keys in the fact table are placed contiguously in a single field to form a **concatenated key** or composite key, the primary purpose of which is to establish the uniqueness of each fact table record. In the physical data base, sequenced integers called **surrogate keys** are often substituted for the primary and foreign keys which has the effect of speeding up queries and leave the data warehouse in a much more flexible position should the key arrangement in the underlying core business systems change. This is unlikely to be a significant factor in Student Record Systems but the designer will want to consider whether or not it is worth while to make the design of the Student Record portion be consistent with the rest of the data warehouse.
SLOWLY CHANGING DIMENSIONS AND MANY-TO-MANY DIMENSIONS

A common abstraction when discussing the dimensional model is to assume that the dimensions are invariant over time. However, in actual practice this is not the case, the descriptions of products and customers, courses and students, are constantly changing. In commercial systems, people change their names, get married and divorced, have more children and move to new addresses. Products continuously evolve, adding and deleting colors, flavors and scents and changing characteristics by make, model and year. In student systems, one of the prime responsibilities of the data warehouse is to correctly represent prior history, particularly for institutional research purposes.

The best practice is not to put all time related attributes in the fact table or to make every dimension time dependent. Instead, the dimensional model exploits the fact that most dimensions are almost constant over time and that the independent dimensional structure can be preserved with only minor additions to capture the time changing nature of most systems. These almost constant dimensions are called slowly changing dimensions.

When changes in the description of products, customers or other sets of attributes are encountered, the data warehouse responds with three main options, referred to as Type 1, Type 2 or Type 3 responses.

Type 1: Overwrite the dimension record with the new values, thereby losing history.
Type 2: Create a new dimensional record using a new value of the surrogate key.
Type 3: Create and “old” field in the dimensional record to store the immediate previous attribute value.

The Type 1 response is used wherever the old value of an attribute has no significance or should be discarded, e.g., the correction of an error. The Type 2 response is the primary technique for accurately tracking a change in an attribute within a dimension. It is used when a true physical change to the dimension entity (like a product, customer, student, course or faculty member) have taken place and it is appropriate to perfectly partition history by the different description. The Type 3 response is used when there is a change that does not perfectly partition history. The Type 3 response is used when there is a change that does not perfectly partition history. There may be a change like the redrawing of admissions district boundaries that is “soft,” i.e., although a change has occurred, it is still possible for queries to be required that assume that the change has not occurred. We may want to track the performance of the Admissions staff within either the old or new district definitions. In this case, history cannot be partitioned disjointly as in the Type 2 response, but both the old and new descriptions are provided in the same record, permitting choices between the two versions of the admissions districts at will.

The instance of Slowly Changing Dimensions are known in advance and their occurrences are chronologically serial, i.e., the time period of one instance doesn’t begin until the time period of the previous instance has ended. However, dimensions may exist that can legitimately have zero, one or many instances per fact table record, the number of which is not knowable in advance. Also, these dimensional instances may be “chronologically overlapping” in that they overlap each other in time. These dimensions are called Many-to-Many Dimensions.

A FEW OBSERVATIONS ON DEVELOPING A STUDENT RECORD DATA MART

The conventional way to develop a design for a data mart is to have lengthy facilitated sessions at which groups of users, assisted by an analyst, develop sets of entities such as customer, product, store and sets of business questions about those entities. The entities and business questions permit the designers to develop a design that can be placed in a dimensional model format, after which development of the data mart can begin.

It can be a lengthy process to gather this information, redact it and edit it into a format helpful for design purposes. While waiting for this process to be organized and implemented there are a few things that users and technological staffs can do to get the design procedure started. A good way to develop a set of data fields for use in the data mart is to place as many computer programs that perform reporting function as possible in a file and parse it, eliminating all text but the field names. Once the list of field names has been compiled they can be divided into fact and dimensional attributes by cardinality, i.e., numeric quantities such as numeric grades, SAT scores, Financial Aid awards and family income data tend to be facts, while textual attributes such as degree, college, major and class level tend to be dimensional attributes.
The normalized entity relationship models of the live files are helpful for suggesting likely dimensional entities and pointing out which attributes are closely correlated. Most Student Record transaction systems place students' personal attributes such as name, birthday, ethnic group, religion and gender in the same file. It is therefore relatively easy to extract these attributes on a nightly basis and place them in the same student attribute or “BioDemo” file in the data mart.

Finally, as the designer begins to think about how to put together the star schemas for the various Student Record constituencies, it should be born in mind that while textual attributes are normally placed in dimension tables and numeric attributes in fact tables, almost all many-to-many relationships, such as the relationship of courses to students and the relationship of Financial Aid awards to Financial Aid recipients, have to be rendered in fact tables as well.

**REGISTRAR STAR SCHEMA**

Registrar Star Schemas are heavily weighted with dimensional attributes; they contain relatively few facts. The usual dimensions of data warehouses built for commercial purposes of customer, product, store, region, etc., provide little guidance for the development of dimensions for a SRDM. However, after parsing the files of production and ad hoc reporting programs, reviewing the structure of the transaction system’s files, consulting with knowledgeable members of the Student Record community and a certain amount of experimentation, the authors arrived at the following set of Registrar related dimensions.

- Time
- Student Matriculation
- Student Bio/Demo (Monster Dimension)
- Institution
- Student Program
- Student Term
- Course
- Cumulative Statistics Student

Most of these dimensions may be applied to all constituencies of the student record system. A brief description of each dimension table follows.

**The Time Dimension**

The Time dimension in a Student Record System is somewhat different from the year, month, week, day sequence of a commercial system. All constituencies, except Financial Aid, operate by Academic Year and Term. The basic units of time for Financial Aid are the Award Year and the Award Period. Table 1 illustrates the relationships between Term, Academic Year and Award Year.

**Student Matriculation Dimension**

The Student Matriculation dimension contains those attributes that describe the student’s current course of study, referred to as the student’s Academic Program Summary, or APS. These attributes include the degree sought by the student, the colleges in which students are enrolled, their majors and minors and the terms in which the course of study began and the terms in which they are expected to end. The Matriculation Table contains one record per student per each course of study, or APS, pursued by the student. Students enrolled in duel degree programs have one record for each program.

**Student Bio/Demo Dimension**

The Student Bio/Demo dimension is analogous to the demographic dimensions in commercial data warehouses in that it contains the characteristics of the university’s “customers,” the students. A standard set of attributes includes name, birth date, gender and ethnic group. In the SRDM it is used among other things to answer queries about diversity and to respond to external survey requests.
Although it is not uncommon in commercial data marts to have a separate address dimension for the customers’ addresses, it is recommended that developers of a SRDM begin by including the addresses in the Student Bio/Demo dimension, forming what is sometime referred to as a “monster” dimension. The reason the Bio/Demo Dimension becomes so large when addresses are included is that students are usually associated with several addresses and individual addresses can serve more than one purpose. The list of possible addresses for each student includes Permanent, Billing, Local, Mailing, Next of kin or Guardian, Grades, and Special and space must be allocated for each address type. Much of this space will be null because few students will have each of these types of addresses.

**Institution Dimension**

The Institution dimension contains the attributes of previous schools with which the college’s students were affiliated. The primary key is a seven digit institution index and the table contains all secondary schools, colleges, corporations, libraries and other institutions with which the college has student related communications. This is a “slowly changing” dimension that requires a type 1 response. It may be updated periodically by a tape issued by the Educational Testing Service.

**Student Term Dimension**

Because data warehousing developed initially in the commercial sector, data modelers have a tendency to seek analogues to Student Record entities in commercial models. The Student Bio/Demo dimension is analogous to Demographics dimensions in commercial applications. Facility dimensions are analogous to Store dimensions and Student Address dimensions are obviously analogous to Customer Addresses dimensions. However, the position of the Student Term dimension in the commercial scheme of things is less obvious. The Term dimension tracks what the student is actually doing in each Term as opposed to the Matriculation dimension which tracks what the student is supposed to be doing.

In the case of the model being developed in this paper, the Student Term dimension tracks the degree, college and major for which students are taking courses in the current Term for up to four possible programs. It also tracks the student’s status as to full-time/part-time, regular enrollment or thesis research, on-campus or study abroad, all of which may change each Term and all of which must be queryable over time.

The best commercial analog to the Student Term dimension is the Customer dimension that tracks the customers’ (students’) “tastes and buying habits” over time. As such, it is a slowly changing dimension requiring a Type 2 response. History is perfectly partitioned by the Term index proxied in the surrogate key that governs each term. The Student Term Table consists of one record per Term per student for each Term included in the SRDM. The four individual programs are treated separately in the following (Student Program) section. If possible, the SRDM should be configured so that users cannot access the Student Term and Student Program dimensions in the same star schema.

**Student Program Dimension**

The Student Program dimension makes available to the user a table of data that includes one record per term, per student, per program in which the student is enrolled. As stated above, it is possible for a student to be enrolled in as many as four distinct programs during each term. Users who wish head counts or rosters of students in a particular program (History majors, for example) can query the system for such without having to know whether the student’s History major is primary or secondary, first or second degree. Note that many of the fields in Student Term Table are repeated in the Student Program Table. This is an example of significant redundancy deliberately being built into the system in order to facilitate different types of queries. If possible, the SRDM should be configured so that users cannot access the Student Program and Student Term dimensions in the same star schema.

**Course Dimension**

The Course dimension contains descriptive material about individual courses such as Course Number (referred to as SECTION_ID), Course Title, Class Room and Instructor. Individual courses may be available in multiple sections, the section numbers being included in the SECTION_ID. The primary key of the Course Table is the concatenation of Term and SECTION_ID (TERMCOURSE) and its grain is one record per Course per Term.
Cumulative Statistics Student

Because of the unusual nature of the grain of the Cumulative Statistics Fact Table (one record per student per academic level obtained; under grad, professional school, etc., discussed below), a dimensional table with a similar time grain is required. The Cumulative Statistics Student dimensional table (CUMSTAT-STUDENT) contains a set of student attributes and a time grain tailored to the purpose of the Current Statistics Star Schema which is to process queries pertaining to students’ overall progress to date at each academic level.

REGISTRAR FACT TABLES

Fact tables are specific to particular constituencies. The following is a list of fact tables for the Registrar, or “Student Record” constituency.

- Student Program Factless
- Student Term Factless
- Attendance
- Current Statistics
- Cumulative Statistics

Student Factless Fact Tables

The Student Factless Fact Tables measure the event of a student being active in a particular Term and Program. They represent the many-to-many relationships between students and all of their dimensions; i.e., many students attend the University for more than one Term, most Terms are attended by more than one student; many students are enrolled in more than one Program, most Programs are taken by more than one student; many students take courses from more than one faculty member, most faculty members teach more than one student, etc. Because these fact tables measure only the event of a student’s attendance in a Program and during a Term, they have only one “fact.” This is a “factless fact,” called STUDENT, which is set to one for all entries in the fact table. This facilitates the aggregation of head counts by any attribute in any dimension.

Because of the difference in grains between the Program Table (one record per Student per Term per Program) and the Term Table (one record per Student per Term), there must be a fact table for each, hence the STUDPROG Fact Table and the STUDTERM Fact Table. See Diagrams REG1 and REG2.

Student Attendance Fact Table

The principal many-to-many relationship in student record data occurs in the relationship between students and the courses that they take. Each student takes many courses; each course is taken by many students. Accordingly, the Attendance Fact Table consists of students and courses with one student and one course in each record. Its grain is one record per student, per term, per course. See Diagram REG3.

Current Statistics Fact Table

The purpose of the Current Statistics Fact Table is to combine in one table those quantifiable measures, pertaining primarily to grades and credit hours, that relate to the current term. The grain of the table is one record per student per term. See Diagram REG4.

Cumulative Statistics Fact Table

The purpose of the Cumulative Statistics Fact Table is to combine in one table those quantifiable measures, pertaining primarily to grades and credit hours, that constitute the students career statistics. The grain of the table is one record per student per Academic Level (Undergraduate, Graduate or Medical, Juris Doctor, etc.) reached by the student. See Diagram REG5.

A FEW CONCLUDING REMARKS

Although only Registrar related data models have been presented, the same principles may be applied to build
data models for the other three student record constituencies. The file structure of query data bases needs to be especially intuitive in Higher Education and the Dimensional Data Model holds the potential to make it so. Parsing existing reporting programs is a good way to develop an initial set of data fields to be included in the data mart and file structures of the entity-relationship core system often provides clues as to the dimensions to be used in the Star Schema. And in particular, determining the grain of the fact table is the first thing to be done in a design of a Star Schema and the grain of the Star Schema’s dimensional tables must be consistent with the grain of the fact table.

Although there is no space here to elaborate, the authors note a few “lessons learned.” Data warehousing is an iterative process, the users must be involved. The Data Warehouse is not a project, it is a process. Its development works best when it is built from the ground up with an emphasis on strategy and a de emphasis on “plan.” Although important in all walks of life, it is especially important in Higher Education to design a development strategy that permits the technical staff to “build a little” followed by the users “getting a little.” Finally, maximizing user participation is crucial. Without it the answer to what is coming to be known in data warehousing as the Field of Dreams Question: If we build it will they come? Is likely to be no.
REFERENCES


<table>
<thead>
<tr>
<th>Obs</th>
<th>YEAR</th>
<th>TERM</th>
<th>ACADEMIC YEAR</th>
<th>AWARD YEAR</th>
</tr>
</thead>
</table>
ATTENDANCE FACT TABLE
DIAGRAM ADM1
ADMIT-TEST-SCORE FACT TABLE

STUDENT-BIODEMO
Dimension
Student-ID_pkey
Student-ID
Info_Release_Flag
Name_Roster
Name_Suffix
Birth_Date
Birth_City
Birth_State
Birth_Country
Gender
Ethnic
Religion
Citizenship
State_of_Origin
Country_of_Origin
Handicap_Type
Visa
Veteran_Code
Veteran_Benefit
Nok_Name
Nok_Relation
E_Mail
Permanent_Street_1
Permanent_Street_2
Permanent_City
Permanent_State
Permanent_Zipcode
Permanent_Country
Permanent_Phone
Local_Street_1
Local_Street_2
Local_City
Local_State
Local_Zipcode
Local_Country
Local_Phone
Billing_Street_1
...
Special_phone
Extract_Datetime

ADMIT-STUDENT
Dimension
Admstud
Student_ID
Appl_College
Appl_Degree
Appl_Major
Appl_Status
Admit_College
Admit_Degree
Admit_Major
Admit_Type
Fin_Aid_Rqstd
Extract_Datetime

TIME
Dimension
Term_pkey
Academic_Year_key
Award_Year_key

ADMIT-TEST-SCORE
Fact
Admstud
Test_Code
Score_Source
Query_Test_Date
Sub_Score_Title
Test_Score
Extract_Datetime

Jan 10, 2000