Reporting Data Kwality Issues

Mike Glasser

Office of Institutional Research
University of Maryland – Baltimore County

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Agenda

- Introduction
- Elements of Data Quality
- Correcting Issues
- Tables and Procedures
- Reports
- Data Quality Firewall
- Questions

UMBC
AN HONORS UNIVERSITY IN MARYLAND

UMBC

University of Maryland - Baltimore County

- Located in suburban Baltimore County, between Baltimore, MD and Washington, DC
- One of the three public research campuses in the University of Maryland System
- 10K undergraduate and 2K graduate students
- 2,200 Employees; 715 full-time faculty
- PeopleSoft, SQL Server 2008 R2
- 9 Pan-Am chess championships

UMBC
AN HONORS UNIVERSITY IN MARYLAND

UMBC RETRIEVERS™

UMBC

Data Warehouse & Reporting

Headcount	IT	IR	Back Office
Management	1	2	3
Data Warehouse	3	2	0
Reporting	3	5	6
FTE	2.5	3	5

- Student Administration module (incl Fin Aid)
- Finance module
- Learn Analytics module (beta)

UMBC
AN HONORS UNIVERSITY IN MARYLAND

UMBC RETRIEVERS

Data Quality

- Accurate
  - Data entry
  - Outdated
- Consistent
  - Records in one table do not match other
- Complete
  - Missing data
- Within Policy
  - Violates a policy or practice
- Reportable
  - OK in system, but cannot be reported

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Accurate Data

- Data Entry
  - Invalid Emplid
  - Invalid Institution
  - Assignment % for Class not 100%
  - Wrong County code for Maryland
- Outdated
  - Wrong contact information
  - Wrong birthdate
  - Student no longer in program

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Consistent Data

- Inconsistent within record or tables
  - US Citizenship, F1 visa
  - Withdrawal code, date before classes
  - Class attribute not on catalog
  - StudentPlan without StudentTerm
  - Enroll Total in Class does not match registrations
  - New Plan not in DW setup table

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Complete Data

- Missing data
  - Plan is missing CIP code
  - Class is missing Instructor
- Missing record
  - Class is missing Instructor
  - Course missing graded component
  - New Plan not in DW setup table

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Within Policy

- Violation of Policy or Practice
  - Double majors for Grad students
  - Duplicate plans
  - 2nd major is Undecided
  - Grad Assistant not enrolled

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Reportable Data

- Cannot report
  - Unknown Gender (IPEDS)
  - Invalid major
- Should not report
  - Non-degree instead of degree seeking
  - Masters instead of Doctoral
  - Class Section changed after freeze

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Remedies

- Meetings or Email
- Transaction system
  - Edit records manually
  - Change business process
  - Change PS data entry (if lucky)
- Data warehouse
  - “Unknown” values (key = -1)
  - Create fake records
    - Bad majors
    - “Two or More” Ethnicity
  - New fields
    - GenderIPEDS
  - Tweak data quality check

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Philosophy

- Fix it in the transaction system
- Prevent / Fix it at data entry
- Fix it as soon as possible
- Fix it before IR census
- Fix it in data warehouse

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Carrot / Stick

- Show them the errors
- Explain the impact
- Exposure to more users
- Data Management Cmte
- Provost
- OK, we will fix it in the DW

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Practice

- Data Quality Team
  - Part of Data Management Cmte
  - Campus commitment to data quality
  - Identify responsible parties
- Data Quality Reporting
  - Identify errors
  - Report errors

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Quality Team

- IR Data Administrator (convener)
- Data Managers/Stewards
  - Registrar
  - Scheduling
  - Financial Aid
  - Undergraduate Admissions
  - Graduate School
  - Finance
  - HR

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Quality Team

- Kickoff meeting to explain philosophy and processes
- Meetings with back office(s) as needed
- Back office report developers
- Write SQL to identify issues
- Write PS Query in transaction system to identify issues
- Develop DW reports with more detail for specific errors

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Quality Reporting

- Identify issue
  - Via meetings, IR, users, back office
- Write SQL
- Nightly procedures
  - Check everything, summarize
- Daily report
  - User subscriptions or on demand
- Fix errors
  - A little prompting

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Tables

- Error Messages
- Daily Errors
- Copy of Yesterday's Errors
- History of Errors
- Exceptions

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Error Messages

- 1 message per error check
- Error Message Number and Text
- Explanation and/or solution
- Module (Admissions, Registrar)
- Table name
- Field name
- Key fields

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Error Messages

- Create table DW.**Data_Quality_Error_Messages** (
  - **Error_Msg_Nbr** int IDENTITY(1, 1),
  - **Error_Msg_Text** varchar(100),
  - **Error_Msg** AS (((Error_Msg_Text + ' [') + CONVERT(varchar,Error_Msg_Nbr)) +']'),
  - **Error_Msg_Severity** varchar(10),
  - **Error_Explanation** varchar(500),
  - **DQ_Module** varchar(3),
  - **Addl_Recipients** varchar(200),
  - **Table_Name** varchar(128),
  - **Field_Name** varchar(128),
  - **Process_Name** varchar(128),
  - **DW_Load_Dttm** datetime,
  - **Key_Fields** varchar(500),
  - **DQ_Sub_Module** varchar(5))

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Daily Errors

- Error Message Number
- Value of the field with error
- Value of fields to identify record
- Date/time
- New error indicator
- Current error indicator

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Daily Errors

- create table DW.**Data_Quality_Daily** (
  - **Error_Msg_Nbr** int,
  - **Fieldvalue** varchar(100),
  - **Key1** varchar(100),
  - **Key2** varchar(100),
  - **Key3** varchar(100),
  - **Key4** varchar(100),
  - **Key5** varchar(60),
  - **Key6** varchar(60),
  - **Key7** varchar(60),
  - **Key8** varchar(60),
  - **Key9** varchar(60),
  - **DW_Load_Dttm** datetime,
  - **Key_Values** varchar(1000),
  - **New_Error_Yn** varchar(1),
  - **Exception_YN** varchar(1),
  - **Current_Yn** varchar(1))

SQL

- SELECT statement to identify the errors
- Has to return the key fields and the field with error (if applicable)
- Convert to INSERT statement for table **Data_Quality_Daily**

- Written by me or back office report developer

UMBC
AN HONORS UNIVERSITY IN MARYLAND

SQL

Academic Plan has invalid CIP code

```
INSERT ( DW_Load_Dttm, Fieldvalue, Error_Msg_Nbr
        , Current_YN, Key1, Key2)
SELECT getdate(),  a.CIP_Code,          180,  'Y'
        a.ACAD_PLAN, a.EFFDT
    FROM iPSSA.Source.PS_ACAD_PLAN_TBL   A
    LEFT JOIN iPSSA.Source.PS_CIP_CODE_TBL B
    WHERE b.CIP_CODE  IS NULL
```

* Removed effective dating logic for simplicity

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Error History

- Summarize daily error messages
- One record per message per day
  - Errors
  - Unique values
  - Exceptions
  - New errors
  - Current errors

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Error History

- create table DW.**Data_Quality_History** (
  - **Error_Msg_Nbr** int,
  - **ETL_Load_Dttm** datetime,
  - **ETL_Date** varchar(10),
  - **Error_Count** int,
  - **Unique_Values_Count** int,
  - **Exception_Count** int,
  - **DW_Load_Dttm** datetime,
  - **New_Error_Count** int,
  - **Current_Count** int)

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Exceptions

- Some errors can be warnings
- Exceptions are on individual case
- Still identified during the checks
- Exceptions are deleted from daily errors after summary, so exceptions can be counted

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Table for Exceptions

- create table DW.**Data_Quality_Exceptions** (
  - **Error_Msg_Nbr** int,
  - **Fieldvalue** varchar(100),
  - **Key1** varchar(100),
  - **Key2** varchar(100),
  - **Key3** varchar(100),
  - **…..**
  - **Key8** varchar(100),
  - **Key9** varchar(100)

Error_Msg_Nbr	Fieldvalue	Key1	Key2	Key3
179	%	SPSE PBC	%	%

Error Message 179 : Academic Plan is missing Degree code

Procedures

- Nightly Data Quality
  - Run Data Quality Checks
  - Cleanup error table
  - Identify errors as new
  - Summarize errors
  - Delete exceptions
- Data Quality Checks SA
  - SQL for individual DQ checks

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Reports

- Microsoft Reporting Services
- Summary report for any date
  - Count of errors, current and new
  - Broken down by module
  - Links to details of error message
- Detail report for last night
  - Which records had the errors
- List of possible error messages
  - With explanation and relevant data

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Summary Report

Error Messages from Oct 9 2012 5:46AM

Data Quality Module	Error Message	Errors	Unique Values	Except	New Errors	Current
AD	**1 messages**	**3**	**1**	**0**	**0**	**3**
AD	Student has invalid Plan in Admissions stack	3	1	0	0	3
CC	**1 messages**	**4**	**4**	**0**	**0**	**4**
CC	Invalid EMPLID	4	4	0	0	4
IR	**2 messages**	**28**	**24**	**0**	**0**	**2**
IR	Academic Plan not found in OIR.DW.UW_Acad_Plan_Tbl	2	2	0	0	2
IR	Class Section changed after Day 10	26	22	0	0	0
SR	**18 messages**	**1290**	**154**	**13**	**0**	**506**
CORE	Academic Plan has an invalid CIP code	1	1	0	0	1
DEG	Degrees in different Plan, same HEGIS	24	5	0	0	0
DEG	More than one degree with same Academic Plan	17	10	0	0	0

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Report of Daily Errors

Data Quality Details for Last Night

Academic Plan has an invalid CIP code [180]

DQ Module : **Student Records (CORE)**
Table Name : **PS_ACAD_PLAN_TBL**
Field with Error : **CIP_Code**
Keys in Table : **Acad_Plan ~ Effdt**

Total Errors : **1**

Error Value	Keys	New	Current
	ACCT UDC~Jan 2 1901 12:00AM	N	Y

- Error Message with Number
- Field with error
- Keys in table
- Value of field with error
- Value of keys to identify record
- Error explanation on hover

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Subscriptions

- Microsoft Reporting Services
- Setup email recipients
  - Can be anyone
  - Don’t need access to report
- Any schedule
  - Recommend daily new, weekly all
- Set report parameters
- Choice of formats
- Link to report online

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Email Summary

Error Messages from Oct 9 2012 5:46AM

Data Quality Module	Error Message	Errors	Unique Values	Except	New Errors	Current
AD	**1 messages**	**3**	**1**	**0**	**0**	**3**
AD	Student has invalid Plan in Admissions stack	3	1	0	0	3
CC	**1 messages**	**4**	**4**	**0**	**0**	**4**
CC	Invalid EMPLID	4	4	0	0	4
IR	**2 messages**	**28**	**24**	**0**	**0**	**2**
IR	Academic Plan not found in OIR.DW.UW_Acad_Plan_Tbl	2	2	0	0	2
IR	Class Section changed after Day 10	26	22	0	0	0
SR	**18 messages**	**1290**	**154**	**13**	**0**	**506**
CORE	Academic Plan has an invalid CIP code	1	1	0	0	1
DEG	Degrees in different Plan, same HEGIS	24	5	0	0	0
DEG	More than one degree with same Academic Plan	17	10	0	0	0

UMBC
AN HONORS UNIVERSITY IN MARYLAND

HR Data Quality

- Process built prior to BbA
- Similar to SA process, but not same tables or procedures
- Email summary sent daily with NEW errors, weekly with ALL
- Email created with SQL, not RS
- Details reported with Crystal
- 180 error checks

UMBC
AN HONORS UNIVERSITY IN MARYLAND

HR Email

Comparing yesterday's errors with today, the following NEW errors were found ...

```
 Count  Error Message
 -----  -------------
     1  Eligible retirement code not found  [151]
     1  EMPL_CLASS inconsistent with EEO6CODE  [113]
     1  Unable to find Benefit_Plan for employee  [81]
     6  EMPLID not found in UM_Person_Info table  [103]

=====================
     9  NEW errors
  1085  Total errors
```

Email produced by Email_New_ETL_Errors_Sp on Oct 9 2012 12:55AM

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Quality Firewall

- Procedures
  - missing data
  - foreign key discrepancies
  - duplicate source keys
- Warnings
  - Usually missing data
  - Loaded as “Unknown” (Key = -1)
- Critical Errors
  - Usually duplicate source keys
  - Only the first is loaded

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Data Quality Firewall

Why don't we use it?

- I already had a system in place for HR data
- I did not know much, if anything, about it
- I did not know XML
- Maybe we could, if time were invested

UMBC
AN HONORS UNIVERSITY IN MARYLAND

UMBC RETRIEVERS™

Recap

- Identify people responsible for data quality in each area
- Back office commitment
- Identify issues and resolutions
- Use DW to capture and report issues
- Report issues to appropriate people
- Impact of data quality
- Accept that some things are wrong only for reporting

UMBC
AN HONORS UNIVERSITY IN MARYLAND

Wrap Up

Any Questions?

Mike Glasser

University of Maryland - Baltimore County

mglasser@umbc.edu (410) 455-3577

Source code is available upon request

UMBC
AN HONORS UNIVERSITY IN MARYLAND