Data Quality 1

MSBO Certification course

Rob Dickinson, MPAAA Executive Director

1

Data Quality 1

- Session Agenda
  - Putting DQ in context
  - Defining our terms
  - Why is quality data so hard to achieve?
  - Finding the problems
  - Fixing the problems
  - Quality Assurance & System Design
- Questions

2

3

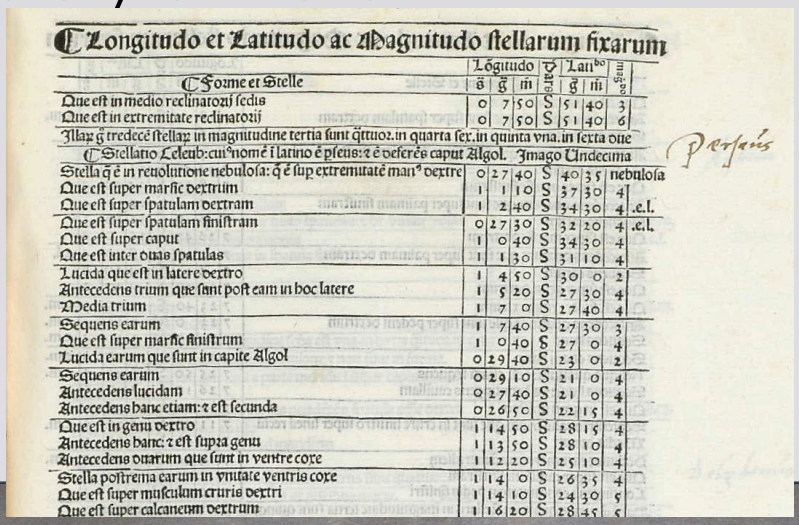
History of Data

- Data – information recorded in an organized fashion for quick lookup and retrieval
- 150 AD – Ptolemy Star Catalog
- List of 1,100 stars, their constellations, brightness, position, etc.

3

4

History of Data



Longitudo et Latitudo ac Magnitudo stellarum fixarum

Forme et Stelle	Longitudo			Pars	Latitudo			Magnitudo
	s	g	m		g	m		
Que est in medio reclinatorij sedis	0	7	50	S	51	40	3	
Que est in extremitate reclinatorij	0	7	50	S	51	40	6	
Illaz g tredecē stellaz in magnitudine tertia sunt ḡttuoz. in quarta ser. in quinta vna. in sexta octe								
Stellatio Celeub: cui nomē i latino ē pseus: z ē deferēs caput Algol. Imago Undecima								
Stella q̄ ē in reuolutione nebulosa: q̄ ē sup extremitatē manꝰ dextre	0	27	40	S	40	35		nebulosa
Que est super marsic dextrum	1	1	10	S	37	30	4	
Que est super spatulam dextram	1	2	40	S	34	30	4	.e.l.
Que est super spatulam finistram	0	27	30	S	32	20	4	.e.l.
Que est super caput	1	0	40	S	34	30	4	
Que est inter duas spatulas	1	1	30	S	31	10	4	
Lucida que est in latere dextro	1	4	50	S	30	0	2	
Antecedens trium que sunt post eam in hoc latere	1	5	20	S	27	30	4	
Media trium	1	7	0	S	27	40	4	
Sequens earum	1	7	40	S	27	30	3	
Que est super marsic finistrum	1	0	40	S	27	0	4	
Lucida earum que sunt in capite Algol	0	29	40	S	23	0	2	
Sequens earum	0	29	10	S	21	0	4	
Antecedens lucidam	0	27	40	S	21	0	4	
Antecedens hanc etiam: z est secunda	0	26	50	S	22	15	4	
Que est in genu dextro	1	14	50	S	28	15	4	
Antecedens hanc: z est supra genu	1	13	50	S	28	10	4	
Antecedens ouarum que sunt in ventre core	1	12	20	S	25	10	4	
Stella postrema earum in vnitate ventris core	1	14	0	S	26	35	4	
Que est super musculum cruris dextri	1	14	10	S	24	30	5	
Que est super calcaneum dextrum	1	16	20	S	28	45	5	

4

5

History of Data

- 1st large database – Untied States Social Security Number system
- Established in November of 1935 – 25 million numbers initially issued.
  - 1st number went to John Sweeney Jr, of New Rochelle, NY
  - Over 1,000 post offices (later SSA offices) had to coordinate issuance of numbers, recording of inc

5

6

History of Data

- SSN – 293-47-2031
- Issued by Toledo OH office in 1971
- Issued to someone who's last name began with 'D'
- 31st person that year whose last name began with 'D' to get a SSN at the Toledo office in 1971
- Identity theft, anyone?

6

7

History of Data

- 1st data issue – Hilda Witcher
- 1936 – EH Ferree Company – New York
- New national contract for Woolworths Dept Store
- Clear box covers (new acetate plastic)
- Wanted something to put in the wallets to catch attention
- SS cards had been going out for around 3 months, lots of newspaper attention

7

8

History of Data

- 1st data issue – Hilda Witcher

8

9

Putting DQ in Context

- Data Quality is one part of larger model – Data Governance
- Data Governance:
  - Policies, processes, and practices that control our data and ensure its quality
  - Hard to see directly, easier by example:

9

10

Putting DQ in Context

- Where most organizations are:
  - Data is defined inconsistently across systems
  - Student data is duplicated
  - Staff time wasted massaging data
  - Fragmented view of students exists
  - Accuracy issues in key data elements
  - Inefficient, leads to 11th hour scramble

10

11

Putting DQ in Context

- The goal is:
  - Key data elements sync across systems
  - Student information is not duplicated
  - Staff spends time analyzing, not verifying
  - Systems show a COMPLETE picture of student
  - Systems report efficiently for all compliance needs
  - Certification deadline is just another day

11

12

Putting DQ in Context

- Not just data
  - How well is staff trained on data definitions?
  - Are field 'owners' known to all?
  - How are staff informed of inevitable changes in these things?
  - Are staff encouraged to analyze data?
  - Does EVERY staff know data privacy rules, and live them?
- All these things add up to Data Governance

12

13

Defining our terms

- Data Quality
  - 2 primary focuses
- Quality Assurance
  - Methods and processes to keep bad data from getting into systems
- Quality control
  - Ways to find and correct bad data once it's in our systems

13

14

Defining our terms

- Pupil Accounting terms
  - FTE – Full Time Equivalency
  - CEPI – Center for Educational Performance and Information
  - MSDS – Michigan Student Data System
  - General Collection – 3 time/year snapshot data collection
  - SRM – Student Record Maintenance
  - EEM – Educational Entity Master

14

15

Defining our terms

- Pupil Accounting terms
  - UIC – Unique Identification Code
  - PIC – Personnel Identification Code
  - SIS/SMS/SRS – Student Information/Management/Record System

15

16

Defining our terms

- Pupil Accounting terms
  - MDE – Michigan Department of Education
  - OAS – Office of Accountability Services
  - Secure Site – District user site maintained by OAS for districts to manage the testing of students

16

17

Defining our terms

- Data Quality
  - Data that is fit for its intended use
- Not "Perfect data"

17

18

Why is it so hard?

- Complex systems becoming interlinked
  - Student records – Special Ed
  - Student records – Food service
  - Student records – Bus Routing
  - Student records – Personnel / HR
  - Student Records – Financial?
  - Student Records – Public portals/Websites

18

19

Why is it so hard?

- Projects are very goal driven, usually compliance driven & punitive (negative feedback only)
  - Fosters attitude "meet THIS requirement, then move on"
- Do you/anyone have time to analyze data, find & fix errors?
- Does your work environment/culture value or invest in data quality?

19

20

Why is it so hard?

- Software Development methods don't emphasize data quality
  - Systems are evaluated on functions/features
  - Look of screen, how it functions
  - Information integrity is not valued as a decision-making criteria
  - If customers don't ask for data integrity, vendors won't build it.

20

21

Why is it so hard?

- Data Quality exercise:
  - Form being handed out
  - Create data input rules for date of birth
  - Ages where warning or errors SHOULD occur
  - Building or District level?
  - Work with your table

21

22

Why is it so hard?

- Data problems are hard to find
  - Data quality evaluated by different systems than capture it
  - Delays in time
  - Personnel, source data no longer available
  - Once problem propagates, much harder to root out.

22

23

Why is it so hard?

- Input fitness
  - Data is usually only made clean to the level needed by the person inputting it.
  - Example - building staff, working with parent
  - No incentive to maintain high data quality
  - Errors only show when data is summarized and integrated – which is usually at time of reporting

23

24

Why is it so hard?

- Do you know all your data's "Intended use"?
  - Data exists in our systems a LONG time
  - Impossible to know ALL intended uses at time of entry
  - Collection systems can't anticipate every future need
  - Reactive legislature adds to the problem

24

25

Who Cares!?!

- Why is quality data important?

25

26

Who Cares!?!

The price of quality data is far lower than the cost of the alternative

26

27

Who Cares!?!

- Costs of bad data
  - Financial
  - Embarrassment
  - Reputation

27

28

Finding the problem

- Who are the flag wavers?
  - Who will know something is wrong?
  - How can they fix it, or raise flag?
- How can YOU find errors?
  - What data can you analyze?
  - What's be best time for analysis?

28

29

Finding the problem

- Flag waivers
  - Parents
    - Parent portal
      - Opportunity to send message on bad data
        - Double edged sword

29

30

Finding the problem

- Flag waivers
  - ETL processes & partners
    - Data errors that occur when files transfer to other software
      - Nightly with special ed system
      - Occasionally with transportation/other systems

30

31

If you want it done right…

- Do it yourself!
  - Compare summary data
    - Do totals by various categories seem close?
      - Need good feel for your data

31

32

Finding the problem

- It can be challenging
  - Student ethnicity – African American
  - CEPI data "010000" or "000100"?
- You shouldn't rely on memory

32

33

Finding the problem

- Data analysis
  - Can you run queries?
  - Download data from your SIS?
  - Can reports be dumped to Excel?
- Beware of privacy issues!

33

34

Deciding to fix the problem

- Can the data be fixed?
- Should the data be fixed?

34

35

Can the data be fixed?

- Is the correction a new value for this field?
  - No – Easy fix
  - Yes – Not so easy

35

36

Can the data be fixed?

- New values for a field
- Mechanical issues
  - Will new value fit nature of field?
  - Will it fit in size of field?
- Policy/process Issue
  - What else will new value affect?

36

37

Can the data be fixed?

- New data values – Effects:
  - All existing reports, queries
  - Any existing error checking
  - Other users of this field
  - Other systems that link at this data

37

38

Should the data be fixed?

- Reasons NOT to fix data
  - Fix has no net benefit
  - Cost greater than benefit
  - Lack of resources
  - Internal need greater than external

38

39

Fixing the problem

- At what level should the fix be executed?
- How should the error be fixed?
- Who should fix the error?

39

40

What level?

- Sometimes, you don't fix the data
  - Internal needs override external
  - Different uses at different levels
- Options:
  - Fix the PROCESS, not the data
  - Create a reporting field

40

41

Create Reporting field

- Internal field used to break out, or consolidate data for reporting
- Place to build in checks
- Usually NOT available to users
- Change data submission routine to use this field

41

42

Fixing with Query

- ALWAYS use 2 step process

1. Run list of errors that shows bad data to be changed as it exists **before** the change
2. Use EXACT same logic to fix error

42

43

Fixing with Query

- Keep list of errors, showing how data WAS before the correction
- ALWAYS list before changing
- The more you can do, the more you can damage!

43

44

Fixing with direct input

- One-off error, single fix
- Run through user interface whenever possible
- Allows any existing error traps to run

44

45

Who should fix the data?

- Users vs Finders
  - Users are the inputters of the data in normal day to day usage
  - Finders are the data staff who collect/report/analyze the data

45

46

Who should fix the data?

- User fix
  - Closest to that particular piece of data
  - 'Owner' for Data Governance
  - Typically, small part of their job
  - Hard to get their attention

46

47

Who should fix the data?

- Finder fix – District level staff
  - Has greatest motivation to make data right
  - Able to concentrate on correction
  - Disconnecting users from their data

47

48

Fixing the data: A special case

Fixing errors in CEPI submissions

- A special case
  - Two approaches
    - Reload
    - Online correction

48

49

Errors in CEPI Submissions

Reload data

- Fix the error in your system, recreate the upload file, re-upload
  - Eliminates synchronization errors
  - Longest process

49

50

Errors in CEPI Submissions

Online correction

- Fix the error in CEPI online system
  - Fast fix
  - Now your source data is different than your reported data

50

51

Errors in CEPI Submissions

What is your transition point?

- First load w/many errors, reload
- At some point in process, switch to online
- Keep synchronization in mind

51

52

Make your life easier

What can you do to make data quality easier?

- Input forms match input screens
- Controls are LIMITED to avoid bad data
- Keep data originators there while you input

52

53

Putting DQ in Context

- Data Quality
  - 2 primary focuses
- Quality Assurance
  - Methods and ways to keep bad data from getting into systems
- Quality control
  - Ways to find and correct bad data once it's in our systems

53

54

Quality Assurance

- Controlling data as it enters your systems
- Important part of system design/installation & maintenance
- 3 areas
  - Data field design
  - Input control functions
  - System modification/customization

54

55

Data Field Design

- Selecting the most appropriate type of field for the data it will hold and assigning properties to that field to limit bad inputting.
- Field Types: Boolean, number, text, date
- Coded fields: Intrinsic, non-intrinsic
- Field Formats: Check boxes, buttons, selection lists, input fields

55

56

Field Types

Boolean

- ONLY 2 values - Yes/No, True/False
- Status (Participant status, Enrolled, Was Absent on Count day)
- Can NEVER hold a 3rd option
- Usually cannot be left blank, or blank is considered one of the values
- Won't allow for any future re-definition

56

Field Types

Number

- Used for values, amounts
- Sometimes used for codes
- Significate digits are important
- Subtypes
  - Integer – 1, 2, 3 (no decimal)
  - Currency – Always 2 digits of decimal
  - Floating Point – No functional limits

57

Field Types

Text

- Used for list of values, string input
- WEAK choice for number only input
- Direct input – Almost impossible to analyze
  - List of options (listbox) gives greater control
- Using text for numbers
  - Allows leading '0', fixed width
  - Only for list of codes

58

59

Field Types

Dates

- Used for inputting dates, sometimes times
- Sometimes stored as number
- Usually built-in error checking for valid dates
- Allows date math
- Formatting for century (3/1/2016 vs 3/1/16)

59

60

Code Fields

- Stores limited list of values
- List determines field type (number, text, etc)
- Good error checking
- Adding & deleting values is a problem
- When creating – Intrinsic vs non-intrinsic
  - Intrinsic – the stored data conveys information
  - Non-intrinsic – stored value has no meaning on its own

60

61

Code Fields

Intrinsic or Non-intrinsic?

UIC

SSN

MSDS Exit codes '19'

MSDS Ethnicity codes '010000'

EEM District codes '41010'

EEM Building Codes '03921'

61

62

Code Fields

Intrinsic codes

- SSN, Gender, Special ed program codes
- Good
  - Easy to understand
  - Built in error checking
  - Can be generated by anyone who knows the rules
- Bad
  - No privacy – allows guessing (identity theft)
  - Needs strong rules
  - Limits possible values
  - Need to know all possible values

62

63

Code Fields

Non-intrinsic codes

- UIC, EEM Building codes, MSDS Exit codes
- Good
  - Not limited by rules
  - Can accommodate growth/change
- Bad
  - Has no value in itself, needs value chart/list
  - Can run into limits (field width)
  - Can only work if there is only 1 place generating values

63

64

Code Fields

Intrinsic or Non-intrinsic?

UIC – *Non-intrinsic*

SSN – *Was intrinsic, changed late 90's*

MSDS Exit codes '19' – *Non-intrinsic*

MSDS Ethnicity codes '010000' - *Intrinsic*

EEM District codes '41010' - *Intrinsic*

EEM Building Codes '03921' - *Non-Intrinsic*

64

65

Field Formats

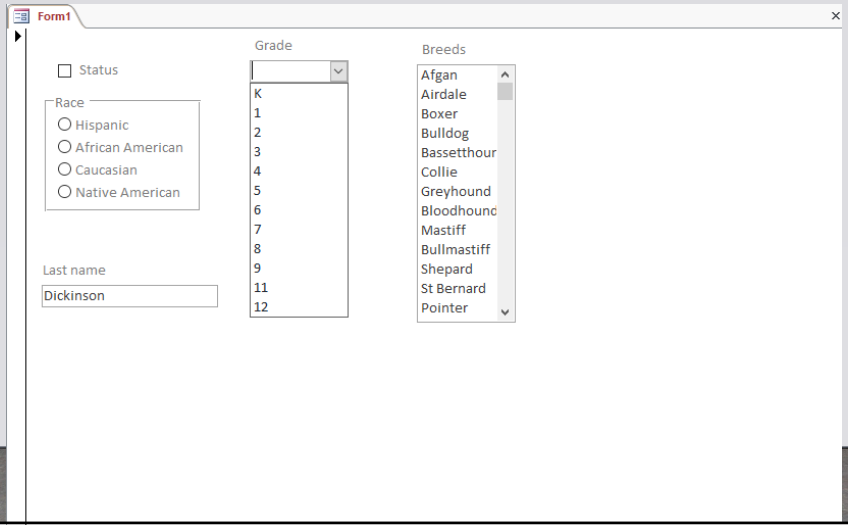
The interface that controls how the data is entered

- Checkboxes, radio buttons
  - Boolean data, 1 choice among very few
- Lists, Dropdown lists
  - List choices available, one or more than 1
- Input box
  - Most freeform, hardest to control input

65

66

Field Formats



66

67

QA Methods

Ways to ensure data is entered into your systems correctly

- Error checking at input
- Training for input staff
- Error checking routines run at regular intervals
- New screens, reports, queries follow same rules

67

68

Error checking at Input

Prevent bad data from getting into the system

- Data Types, field formats
- Error checking rules behind the field
- Make it difficult to allow non-standard data to be input
  - Can't make it so hard that it is ignored
  - 'Are You Sure?'

68

69

Training for Input Staff

Make sure staff entering data is aware of its importance

- Initial training
  - Bring new staff up to speed
  - Familiar with systems
- Recurring training
  - Letting everyone know what's new, changed
  - Reminders on problem areas

69

70

Error checking routines

Frequently run reports/queries designed to find errors soon after input

- Find and fix before it is used, propagated to other systems
- Nightly, over weekend, end of attendance period
- Can be system report, email, faxed, etc.
- Do you fix, or do they?
- Balance of finding errors vs overwhelming users

70

71

Error checking routines

New screens or reports or queries MUST follow same QA rules to prevent 'trapdoor' errors

- New screens – 'All in one' or audit screens
- Reports – Do report generators follow security rules?
- Queries – double edged sword
  - Easy to run, change data quickly
  - Usually avoids all the user input rules
  - Can fix - or break - large amounts of data very fast

71

72

Getting Help

- Pupil Accountant
  - Closest, knows your data best
- ISD Auditor
  - Knows the rules and regulations
  - Does your Pupil Accountant want to be a liaison?

72

Getting Help

- CEPI Helpdesk
  - (517) 335-0505, Option 3
  - cepi@michigan.gov
- MPAAA
  - Rob@mpaaa.org
  - (517) 853-1413